DOCUMENT RESUME

ED 140 817 IR 004 914

AUTHOR	McEwen, Hazel E., Ed.
TITLE	Management of Data Elements in Information Processing. Proceedings of a Symposium (2nd, Gaithersburg, Maryland, October 23-24, 1975).
INSTITUTION	National Bureau of Standards (DOC), Washington, D.C. Inst. for Computer Sciences and Technology.
SPONS AGENCY	American National Standards Inst., Inc., New York, N.Y.
REPORT NO	PB 249530
PUB DATE	Oct 75
NOTE	275p.; For related document, see ED 093 370 ; Not available in hard copy due to print quality of original
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (SD Cat. No. C13, MF $2.25, HC $9.25)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	*Codification; Computer Science; Data Bases; *Electronic Data Processing; *Information Processing; Information Systems; *Management; Medicine; Merchandise Information; On Line Systems; Post Secondary Education; *Standards; Symposia; Telecommunication
IDENTIFIERS	American Standard Code Information Interchange; Data Element Dictionary Directory

ABSTRACT

Continuing technological advances in computers and communications make possible the integration of data systems and the exchange of data among them on an expanding scale. However, the full effect of these advances cannot be realized unless the need for uniform understanding of the common information (data elements) and their expression in data systems is recognized and a means provided to effectively manage this information. The increasing interrelationships among the data systems of Federal, State and local governments, and with industry and the public add emphasis and dimension to the need for the improved management of data elements in information processing. These Proceedings are for the second Symposium on the Management of Data Elements in Information Processing held at the National Bureau of Standards on 1975 October 23-24. Over 300 representatives of Federal and State governments, industry and universities from 29 states, from Japan, and the United Kingdom were in attendance. Twenty-nine speakers discussed the role of the data manager, communications needs for data standards, data element directories, standard codes for character and control, use of check characters, data elements in bibliographic data bases, product coding, coding for clinical medicine, human factors, data resource management, data base management systems, and other subjects related to data standardization and data management efforts. (Author)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED140817

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PB-249-530

Management of Data Elements in Information Processing

Proceedings of a Second Symposium Sponsored by the American National Standards Institute and by The National Bureau of Standards

1975 October 23-24
NBS, Gaithersburg, Maryland

BEST COPY AVAILABLE

Hazel E. McEwen, Editor

Institute for Computer Sciences and Technology
National Bureau of Standards
Washington, D.C. 20234

U.S. DEPARTMENT OF COMMERCE, Elliot L. Richardson, *Secretary*
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Acting Director*

IR004914

Table of Contents[1]

Page

Introduction to the Program of the Second National Symposium on The Management of Data Elements in Information Processing ix
David V. Savidge, Program Chairman

On-Line Tactical Data Inputting: Research in Operator Training and Performance 1
Irving Alderman, Ph.D.

"Turning the Corner" on MIS, A Proposed Program of Data Standards in Post-Secondary Education 9
Donald R. Arnold, Ph.D.

ASCII - The Data Alphabet That Will Endure 17
Robert W. Bemer

Techniques in Developing Standard Procedures for Data Editing 23
George W. Covill

An Adaptive File Management Systems 45
Dennis L. Dance and Udo W. Pooch (Given by Dance)

A Focus on the Role of the Data Manager 57
Ruth M. Davis, Ph.D.

A Proposed Standard Routine for Generating Proposed Standard Check Characters 61
Paul-Andre Desjardins

Methodology for Development of Standard Data Elements within Multiple Public Agencies 69
L. D. England, S. L. Eberle, B. H. Schiff and A. S. Huffman
(Given by England)

The Role of the Internal Auditor in Data Management 77
Richard H. Fahnline

Semantic Coding and Data Element Characterization in Medical Computing 83
E. R. Gabrieli, M.D., F.C.A.P.

Principles and Concepts of Data Resource Management System Development 91
Aaron Hochman

---

[1] Papers are sequenced by name of author.

The Design of Data Elements: A Data Base Perspective 99
M. A. Huffenberger

A Challenging Aspect of Word Processing 113
Victor G. Kehler

Data Element Lexicon Needs A New Home 119
Richard J. Kirkbride

Check Characters and the "Self-Checking String" --
What, Where, Why, When and How 121
J. R. Kraska, J. R. Nelson and E. Hellerman
(Given by Kraska and Hellerman)

The Standards Implications of the Developing Interrelationships Between On-Line Bibliographic Retrieval,
Data Manipulation and Micrographics Display 139
Robert M. Landau

Product Coding-One Number from Maker to User 147
John T. Langan

Development of a Data Dictionary/Directory Using a Data
Base Management System 151
E. K. C. Lee and E. Y. S. Lee

Systems Design Considerations for the U.S. Army Materiel
Command (AMC) Data Element Dictionary/Directory System 163
Fernando Puente

A Data Element Directory for a State Motor Vehicles Agency 169
John Roberts

An Integrated Dictionary for Systems and Data Components 191
Curg Shields

An Information Documentation Language: A Framework for
Deriving Information from Data 195
William M. Taggart, Jr.

International Standards for Data Transmission 211
V. N. Vaughan, Jr.

An Information Management View of Data Management 221
Marvin G. Wallis

Data Standardization 231
Harry S. White, Jr.

Data Element Analysis and Use of a Relational Data Base Structure for Mapping Bibliographic and Numeric Data Bases 237
Martha E. Williams, Scott Preece and Sandra Rouse
(Given by Williams)

Status of the Army Materiel Command's Progression from Reports Control to Data Element Management 253
Edith F. Young

Appendix A - Participants 257

FOREWORD

The National Bureau of Standards has been active in the field of computer standards since the early 1960's. In 1973, the President by Executive Order assigned primary responsibilities for data standards to the Secretary of Commerce. NBS has been assigned the responsibility by the Secretary of Commerce for providing administrative and technical support for the Federal-wide data standards program (Part 6 of Title 15 of the Code of Federal Regulations).

The Symposium on the Management of Data Elements in Information Processing was held at the Gaithersburg Laboratories of NBS on October 23 and 24, 1975 and was jointly sponsored by NBS and the American National Standards Institute Committee X3L8 on the Standardization of the Representation of Data Elements. David V. Savidge served as the general chairperson of the symposium and Hazel E. McEwen was the arrangements chairperson. Three hundred and twenty-eight professionals attended from twenty-nine states and two foreign countries. Twenty-six technical papers were summarized in the oral presentations given. The Bureau is pleased to have the opportunity of making the full text of the papers presented at the symposium available in these proceedings. We feel that these are representative of the state-of-the-art of current data management practices and should provide a useful base for further activities in this important field.

However, it must be stressed that the responsibility for the content of the papers provided in these proceedings rests with the individual authors and their organizations and do not reflect endorsement by the National Bureau of Standards.

Harry S. White, Jr.
Associate Director for ADP Standards
Institute for Computer Sciences and
Technology, NBS

INDEX OF
KEY WORDS AND PHRASES
TO
AUTHORS' NAMES

ADAPTIVE FILE MANAGEMENT - Dance/Pooch
ALPHABET - Bemer
AMERICAN NATIONAL STANDARDS - White
AMERICAN STANDARD CODE FOR INFORMATION INTERCHANGE (ASCII) - Bemer
APPLICATIONS - Hochman
ARMY MATERIEL COMMAND (AMC) - Young
ARTIFICIAL INTELLIGENCE - Gabrieli

BIBLIOGRAPHIC INFORMATION SYSTEMS - Landau

CCITT (abbrev. for French words meaning Telegraph and Telephone International Consultative Committee) - Vaughan
CENTRAL SYSTEM DESIGN AGENCY (CSDA) - Puente
CHANGE CONTROL - Hochman
CHARACTER - Bemer
CHECK CHARACTER(S) - Desjardins; Kraska/Nelson/Hellerman
CHECK DIGIT - Desjardins
CLASSIFICATION - Kraska/Nelson/Hellerman
COMMON BUSINESS ORIENTED LANGUAGE (COBOL) ROUTINE - Desjardins
CODE - Bemer; Desjardins; Vaughan
CODING - Gabrieli
COGNITIVE MEMORY - Gabrieli
COMMODITY COMMAND STANDARD SYSTEM (CCSS) - Puente
COMMUNICATIONS - Kehler
COMPUTER FACILITY MANAGER - Davis
COMPUTER TECHNOLOGY - Alderman
COMPUTEROUTPUTMICROFORM (COM) - Kehler
CONFIDENTIALITY - Gabrieli
COPY LIBRARY - Puente
COST/BENEFIT - Kraska/Nelson/Hellerman

DATA - Landau; Shields; White
DATA BASE - Huffenberger; Williams/Preece/Rouse
DATA BASE CONCEPT - Huffenberger
DATA BASE MANAGEMENT - Davis; Lee/Lee
DATA BASE MANAGEMENT SYSTEM (DBMS) - Huffenberger; Lee/Lee
DATA BASE MANAGER (DBM) - Huffenberger
DATA CHARACTERIZATION - Gabrieli
DATA DEFINITION - Lee/Lee; Taggart
DATA DICTIONARY - Lee/Lee; Shields
DATA DICTIONARY/DIRECTORY (DD/D) - Huffenberger
DATA DIRECTORY - Lee/Lee; Puente
DATA ELEMENT(S) - Alderman; Huffenberger; Kirkbride; Lee/Lee; Roberts; Taggart; White
DATA ELEMENT CHARACTERISTICS - Young
DATA ELEMENT DESCRIPTIONS - Roberts
DATA ELEMENT DICTIONARY (DED) - England/Eberle/Schiff/Huffman; Puente; Roberts; Young
DATA ELEMENT LEXICON (DELEX) SYSTEM - Kirkbride
DATA ELEMENT MANAGEMENT - Wallis; Young
DATA ELEMENT MANAGEMENT BASE FILES - Young

DATA ELEMENT MAPPING - Williams/Preece/Rouse
DATA ELEMENT MATRIX ANALYSIS - Young
DATA ELEMENT STANDARDIZATION - England/Eberle/Schiff/Huffman; Hochman; Young
DATA ENTRY - Alderman
DATA FIELDS - Hochman
DATA INDEPENDENCE - Huffenberger
DATA INPUTTING - Alderman
DATA INTEGRITY - Huffenberger
DATA INTERCHANGE - Taggart
DATA ITEM - Taggart
DATA KEY VALIDATION - Kraska/Nelson/Hellerman
DATA MANAGEMENT - Fahnline; Shields; Wallis; White
DATA MANAGER - Davis
DATA RELATIONSHIP - Hochman
DATA REPRESENTATION - Hochman
DATA RESOURCE MANAGEMENT - Hochman
DATA STANDARDIZATION - White
DATA STANDARDS - Arnold; Taggart; White
DATA STANDARDS IN PUBLIC AGENCIES - England/Eberle/Schiff/Huffman
DATA STRUCTURE - Lee/Lee
DATA SYSTEM - Lee/Lee
DATA TRANSMISSION - Vaughan
DECISION MAKING - Taggart
DEFENSE COMMUNICATIONS AGENCY - Kirkbride
DEFINITION INFORMATION - Roberts
DOCUMENTATION - Covill; Huffenberger
DOCUMENTS - Hochman

ECONOMICS - Landau
EFFECTIVENESS - Taggart
EFFICIENCY - Kraska/Nelson/Hellerman
ERROR-DETECTING - Desjardins
EXPLANATORY TEXT - Roberts

FEDERAL GENERAL STANDARDS - White
FEDERAL INFORMATION PROCESSING STANDARDS - White
FEDERAL PROGRAM STANDARDS - White
FEDERAL STANDARD - Lee/Lee
FILE GUIDE - Puente
FILE WORKING SET - Dance/Pooch
FILES - Hochman
FILES INVENTORY - Roberts
FORMATS - Hochman
FORMS - Hochman
FORTRAN - Kraska/Nelson/Hellerman

HIERARCHIAL FILE STRUCTURE - Dance/Pooch
HIGHER EDUCATION GENERAL INFORMATION SURVEY (HEGIS) - Arnold
HUMAN FACTORS - Alderman
HUMAN PERFORMANCE - Alderman
HUMAN RELIABILITY - Alderman

IDENTIFICATION INFORMATION - Roberts
INFORMATION - Shields
INFORMATION COMPOSITION - Kehler
INFORMATION DEFINITION - Taggart
INFORMATION DISPOSITION - Kehler
INFORMATION DISTRIBUTION - Kehler

INFORMATION DOCUMENTATION LANGUAGE - Taggart
INFORMATION ELEMENT - Taggart
INFORMATION FLOW - Wallis
INFORMATION INTERCHANGE - Taggart
INFORMATION ITEM - Taggart
INFORMATION MANAGEMENT - Wallis
INFORMATION ORGANIZATION - Landau
INFORMATION REPRODUCTION - Kehler
INFORMATION RETRIEVAL - Kehler
INFORMATION STANDARDS - Taggart
INFORMATION STORAGE - Kehler
INFORMATION SYNTAX - Taggart
INFORMATION SYSTEM - Lee/Lee; Wallis; White
INFORMATION SYSTEMS - Shields
INFORMATION TECHNOLOGY - Davis
INFORMATION TRANSFER - Kraska/Nelson/Hellerman
INTERFACE - Vaughan
INTERNAL AUDIT - Fahnline
INTERNATIONAL ALPHABET NO. 5 - Vaughan
INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO) - Bemer
INTERNATIONAL STANDARDS - White
INTERNATIONAL TELECOMMUNICATIONS UNION (ITU) RECOMMENDATIONS - Vaughan

KEY ERROR TYPES - Kraska/Nelson/Hellerman
KEY WORD IN CONTEXT (KWIC) - Puente
KEYBOARDS - Kehler

LANGUAGE - Landau
LOGISTICS - Puente

MAN-MACHINE INTERFACE - Gabrieli
MANAGEMENT CONTROLS - Fahnline
MANAGEMENT ENTITIES - Hochman
MANAGEMENT INFORMATION SYSTEMS - Arnold
MANAGEMENT PRINCIPLES - Hochman
MANUAL OF DATA ELEMENTS - England/Eberle/Schiff/Huffman
MASTER FILE CONTROL - Puente
MEDICAL CODING - Gabrieli
MEDICAL COMPUTING - Gabrieli
MEDICAL INFORMATION SYSTEMS - Gabrieli
MEDICAL LEXICON - Gabrieli
MEDICAL TAXONOMY - Gabrieli
MEDICAL TERMINOLOGY -Gabrieli
METHODOLOGY - Shields
MICROELEMENT - Williams/Preece/Rouse
MICROFORM - Kehler
MICROGRAPHICS - Landau

NATIONAL CENTER FOR HIGHER EDUCATION STATISTICS (NCES) - Arnold
NUMERIC DATA SYSTEMS - Landau

ON-LINE SYSTEMS - Landau

PHOTOTYPESETTING - Kehler
PLANS - Hochman
PLAYSCRIPT - Covill

POST-SECONDARY EDUCATION - Arnold
POWER - Kraska/Nelson/Hellerman
PRIVACY - Davis; Gabrieli
PRIVACY ACT - Davis
PROCEDURE - Covill
PROCESSES - Hochman
PROGRAMMING MNEMONICS - Puente
PROGRAMMING VIABILITY - Huffenberger
PROGRAMS - Hochman
PUBLISHING - Kehler

RANDOM ERROR - Desjardins
RECORDS - Hochman
RELATIONAL DATA BASE STRUCTURE - Williams/Preece/Rouse
REPORTS - Hochman
REPORTS CONTROL - Young
REPRESENTATIONS OF DATA ELEMENTS - Taggart
RETRIEVAL - Landau
RIGHT-TO-PRIVACY - Gabrieli

SELF-CHECKING - Desjardins
SELF-CHECKING STRING COMPOSITION - Kraska/Nelson/Hellerman
SELF-ORGANIZING DATA SETS - Dance/Pooch
SOFTWARE - Kraska/Nelson/Hellerman
STANDARD(S) - Covill; Desjardins; Landau; Vaughan
STANDARD DISTRIBUTION FORMAT (SDF) - Huffenberger
STANDARD FILE FORMAT (SFF) - Huffenberger
STANDARD SYSTEMS - Fahnline
STANDARDIZATION - Lee/Lee
STANDARDIZATION METHODS - England/Eberle/Schiff/Huffman
STANDARDIZED MEDICAL NOMENCLATURE - Gabrieli
STATE DATA ELEMENTS AND REPRESENTATIONS - England/Eberle/Schiff/Huffman
STATISTICS - Huffenberger
SUBSTRUCTURE SEARCHING - Williams/Preece/Rouse
SWITCHED NETWORK - Vaughan
SYMBOL - Bemer
SYSTEM 2000 - Lee/Lee
SYSTEM CONCEPTS - Hochman
SYSTEM CONTROL AND DOCUMENTATION (SYSCAD) - Puente
SYSTEM PERFORMANCE - Alderman
SYSTEM RELEASE MANAGEMENT - Puente
SYSTEMS - Hochman
SYSTEMS DICTIONARY - Shields

TELECOMMUNICATIONS - Vaughan
TEXT EDITING - Kehler
TRANSACTIONS INVENTORY - Roberts
TRANSCRIPTION ERROR - Desjardins
TRANSPOSITION ERROR - Desjardins
TYPESETTING - Kehler

USER INTERFACE - Landau

WORD PROCESSING - Kehler
WORD REFERENCE - Roberts

Introduction to the Program of the Second National Symposium on the Management of Data Elements in Information Processing

David V. Savidge, Program Chairman
Special Assistant to the President
COMNET
5185 MacArthur Blvd.
Washington, D.C. 20016

Alexander the Great wept because there were no more worlds to conquer. Let's change that to read: because he didn't see any other worlds. If he had seen some, he could have conquered them.

The proponents of standard representations for data elements are almost like Alexander. Standard representations are evolving for the precise entities of time, place, persons, and things. In these areas a picture or a thousand words may be identified by a single symbol. What next?

First, have the requirements of the precise entities been fully met? Colloquialisms are developing very rapidly. Complaints are heard in the scientific community that mathematicians cannot communicate with those in other disciplines because of their language. No one else knows it. On the other hand, benefits of these trends are being reaped in the field of chemistry. As the result of the Standard representations of chemical compounds and their attributes, chemists are able to develop products which have not yet been found in nature. Many scientists believe that if certain properties are required in a substance, chemists can produce substances with those properties. Could this type of synthesis be accomplished in mathematics or other disciplines if it is needed?

Secondly, to what extent can standard representations prove useful in areas of imprecision, where entities are abstract, such as concepts, or where entities are constantly changing? There is a common requirement in areas of precision and imprecision -- communication. However large or small the community of interest, there must be agreement within that community as to the representations used in communications. As complex as our society has become and as specialized as the components of that society are, there always will be the requirement to improve the quality of communications of precise and imprecise information.

This Second National Symposium on the Management of Data Elements, in addition to including topics which were not included in the first one, includes papers which report progress on projects identified in the first symposium and some which were started afterwards. The primary purpose of the symposium is to stimulate thought and to increase the flow of ideas between people with the same kinds of problems. We believe you will be able to relate to many of the situations discussed at this symposium.

Tactical Data Inputting: Research in Operator Performance and Training

Irving N. Alderman

U. S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, Virginia 22209

The Army development of battlefield information systems established a need for a human factors research program directed toward improved integration of human tasks in the system operations. This paper reviews selected studies on the data entry process. The allocation of functions/tasks for data entry and the development of job and training aids are related to human performance and to the effectiveness of the system as an aid to the user.

Key words: Computer technology; data elements; data entry; data inputting; human factors; human performance; human reliability; system performance.

1. Introduction

The U.S. Army program for research and development of military command and control systems includes a human factors effort directed towards improving the effectiveness of battlefield information systems. The primary function of these systems will be to aid the commander and his staff in tactical decision making by selectively automating the currently tedious and time consuming tasks associated with the collection, processing and display of information. These systems may be conceptualized as information systems to support the allocation of resources in a combat environment. The data input subsystem is a critical component of the total system. The human performance (time and error measures) associated with this subsystem impact on the quality, completeness and timeliness of subsequent information processing, assimilation and decision making in the highly dynamic combat environments encountered in modern warfare. This crucial role of the human in the system defines human performance levels which must be retained in a variety of performance degrading work environments (e.g., climate, noise, stress) and tactical situations, if the effectiveness of the system is to be preserved.

The ARI research and development program in Command Systems began as a concept in which field evaluations and laboratory studies form an interactive enterprise. Field evaluations to provide an opportunity to identify

---

The views expressed in this paper are those of the author and not necessarily those of the U.S. Army Research Institute for the Behavioral and Social Sciences. Reproduction in whole or in part is permitted for any purpose of the United States Government.

and orient research questions derived from operations under simulated combat conditions; and laboratory experiments to provide the ability to isolate and control relevant variables and develop potential solutions which can be validated in field evaluations. The requirement to support the design verification of an experimental automated Tactical Operations System (TOS) provided a unique opportunity to explore this program concept.

This paper describes the field and laboratory studies in this program concerned with the data inputting process and the techniques that have been developed to enhance system effectiveness through improved human performance.

2 The System and its Operations

The experimental Tactical Operations System (TOS) was conceptualized as a central data store and computer facility which accepted data inputs and responded to requests for data processing with outputs to the remote users. Three basic types of equipment comprised the system: User Input/Output Device (UIOD), Remote Station Data Terminal (RSDT) and the Central Computing Center (CCC). The UIOD provided the remote users with communications to the system and other users. The RSDT served as an intermediate message processor/transmitter between several UIOD's and the CCC. The CCC contained the data store, processed incoming messages, manipulated the data, and output messages to the RSDT's for the users.

Communication with the CCC was via the UIOD which provided the immediate interface between the user and the system. This device consisted of a desk top terminal, containing a CRT with associated keyboard, and a teletypewriter.

Users of the system employed two general types of messages: "operational" to enter and retrieve information and "relay" for communication between users independent of the CCC. The operational messages were of four basic types: data (enter, change, add, or delete), query (request data base search), special process request (data manipulation), and standing request for information (automatic dissemination). Figure 1 represents schematicly the five critical human functions associated with system operations. These functions may be summarized as follows:

1. Screen the incoming data for pertinence, credibility, impact, and priority.

2. Transform the raw data into computer-acceptable form;

3. Input the transformed data into the computer via an entry device;

4. Assimilate displayed data and

5. Decide on a course of action based on the available information.

This paper is concerned primarily with the transform and input functions. Subsequent system operations are not independent; delays and errors introduced during operation affect subsequent operations and the overall system performance.

Assuming messages originate in free text form as they do in a manual system, the transform operation, as shown in figure 2, includes selection of the proper format for input of the data in the message and conversion of the data into computer acceptable codes. Entry of the completed message formats into the system is by use of the keyboard in the UIOD. This process is shown in figure 2. The preceding description of the program concept and its development as well as the description of the experimental TOS and its operations were abstracted from Baker [1][1].

3. Research in Data Inputting

An early study by Baker, Mace and McKendry [2] compared a job aid for the format selection task with a "menu" type listing of the message types and their reference codes. The job aid consisted of a table of reference codes with the various message subjects (position report, planned action, etc.) shown in the table rows and the message action (modify data, query, etc.) in the columns. Two groups of participants with TOS experience were formed; one group used the aid and the other used the menu to select the appropriate format reference code for a set of message/situation descriptions. There were no reliable systematic differences in the performance of the two groups due to the job aid. However, the data provided a baseline for assessing the impact of this task on the system. The average error was 22% with a mean time to select a format of 50 secs.; suggesting that one message in five will be incorrectly entered and, hence, not available for computer retrieval or processing. This error rate is significantly higher than the system designers' predicted rate.

Format selection is only the initial task in the data entry process (see fig. 2) which also includes input of the formats. During high message traffic periods when message arrival rates will exceed one per minute, a queue may be expected with the attendant delays and degradation in the timeliness of the information. The need to improve performance of the format selection task was apparent from the Baker, Mace and McKendry study [2] and the following study was conducted to explore another potential aid to reduce input time.

Reference codes used in the experimental TOS consisted of two letters indicating a message type and a number to identify an action code (e.g., LL#). An alternative set of codes based on four letters (LLLL), usually an acronym of the message format title and the action name (e.g., add, delete, query, etc.) was recommended. Nystrom and Gividen [3] compared the ease or difficulty of learning the two sets of message format-action codes. Comparison of the error rates showed a reliable difference in learning the codes, i.e., the error rate for learning the LLLL codes was less than half that for learning the LL# codes (13% vs 29% errors for enlisted men; 11% vs 21% for officers working on one sample list, and 7% vs 15% for officers working on another sample list). In addition, the time required to meet a learning criteria for the LLLL groups was approximately 60% of the time required for the LL# groups. Thus, the LLLL code was learned in less time and with fewer errors than the LL# code. A detailed analysis of the error rates suggested several revisions in the LLLL code that have the potential to reduce the error rate to 5% or less.

A study by Strub [4] compared performance obtained using two modes of format preparation (off-line vs on-line) with and without verification. On-line entry used the terminal for preparation; off-line required preparation of a paper format prior to entry into the terminal. In the verified condition, two operators compared their completed formats, resolved any discrepancies between the two, and entered the formats individually. The unverified condition omitted these steps, i.e., the operators performed the input operation without an error check. A fifth control group of operators prepared formats using procedures similar to those employed in the experimental TOS, i.e., formats were prepared on paper by one operator and transcribed onto the terminal by another operator. Significantly fewer errors were obtained in the on-line preparation (11.2%) than in the off-line preparation (14.8%). There were no reliable differences between the five conditions in speed of input. Errors were reduced by approximately one-third by the verification procedure (15.7% vs 10.3%) but approximately one-third more time (4.98 min. vs 6.81 min.) was required for verification. Either procedure was superior to the inputting process used in the experimental TOS. From these findings, Strub [2] recommended on-line inputting to reduce errors and noted that although the verification procedure reduced errors, the tradeoff in time and manpower must be considered.

---

[1] Figures in brackets indicate the literature references at the end of this paper.

In a subsequent study by Strub [5], computer assistance was explored as an aid to the data inputting process. Two types of aiding were used: entry information, instructions and legal entries presented on the CRT versis the same material in looseleaf notebooks was compared using two types of format: full, in which the complete format is provided with blanks to be entered and a partial format containing only those entries that were previously selected by checklist. A fifth control or baseline condition, provided the operator with a reference notebook and a blank CRT for input. There were no significant differences due to either the computer aid or the format. However, the time required under the control condition was significantly longer than either of the comparable full format or checklist methods. A detailed analysis of the input errors was conducted to categorize the errors by type and evaluate the effectiveness of automated edit and validate routines to detect errors (categorized as: commission, omission, glossary, category, abbreviation, or typographic). Commission, omission and typographic errors were identified as computer detectable. Based on the Strub data [4], these three categories accounted for 20% of the errors; the remainder or 80% of the errors would be undetected. The need to reduce the undetected error rate provided the impetus for development of techniques to reduce their probability. Two approaches to this problem are (1) the identification and reduction of the sources of error and (2) the development of training techniques to improve operator proficiency.

Nawrocki, Strub and Cecil [6] developed a technique for the analysis of errors based on comparing the entered message formats with formats having perfect accuracy. Mismatches or discrepancies between corresponding entries being defined as errors. Errors were then categorized (with reference to the free text messages) according to the type of failure in the inputting process (e.g., omission, commission) into a set of mutually exclusive and exhaustive categories. This technique for error analysis was first applied to data obtained in the first study by Strub [4]. Examination of the mean error rates revealed that two of the error categories, omission and typographic, accounted for 60% of the errors; this was consistent across the five system-input procedures. Assuming the system consequences of each error category are similar, the greatest improvement in system performance would be realized by reducing the error rates in these two categories. Omission errors are due to the operator's omission of some of the information in the messages; typographic errors are associated with copying the paper formats and with keying errors. Consequently, a priori, the on-line, verified procedure would be expected to have fewer errors in these categories than the other four inputting procedures used in the study, an expectation supported by the error rates for the categories in each of the five inputting procedures. The technique was also applied to the data from the subsequent study by Strub [5], and its applicability to other man-computer systems involving similar human processes was demonstrated. Thus, operator errors are susceptible to reduction by identification of the causal process and the use of human factors techniques to improve system operations.

Another approach to improved human performance is the development of software/courseware training modules embedded within the operating system to both train operators and maintain a high level of proficiency once trained. The effectiveness of embedded training programs depends on the development of optimized training strategies which are responsive to the student's history, current skill level, etc., and satisfy the necessary conditions for learning, e.g., feedback of errors. A study currently in progress by Alderman and Gade [7] is designed to compare the effects of selective feedback on performance during training and simulated operations. Five conditions were defined by the type of computer feedback provided to the student when an error occurs. The five types of feedback are:

1. No Feedback
2. Minimum Feedback - error message only
3. System Feedback - error message, correction required.
4. Remedial Feedback - error message, correction required, correct entry.
5. Credit Feedback - error message, correction required, correct entry, automatic entry.

The error message informs the student that his last entry was in error; correction required indicates that erroneous entries must be corrected before proceeding; correct entry provides the student with both his erroneous entry and the correct entry. For automatic entry the computer enters certain data elements after the student has satisfied a learning criteria for that data element. Errors are defined as incorrect or illegal entries in all conditions except system feedback in which errors are the illegal entries detectable by edit and validation routines. Performance measures (time and errors) are being obtained as participants train with either minimum, system, remedial or credit feedback and, in a subsequent session, simulate operations with either no feedback or system feedback. Analysis of this data will provide information concerning differences in performance attributable to different types of feedback.

In parallel with the research described above, Baker [8] describes a model of a generalized information system as a tool for evaluating the effects of human performance on overall system effectiveness using various mixes of equipment, personnel and procedures. A computer simulation (MANMOD) derived from this model was developed to simulate system operations based on three dimensions: data flow and processing; human task analysis for each event in the data flow; and a source of variation, e.g., skill level. These dimensions are specified by program inputs and provide considerable flexibility in structuring various conditions of the system being simulated. MANMOD focuses on the performance of data-inputting personnel in an experimental TOS, i.e., an action officer to screen and transform messages and a UIOD operator to enter the data. Personnel characteristics (e.g., speed, precision, level of aspiration, stress threshold and fatigue) are entered as initial values and modified in response to simulated events (e.g., development of queues and errors). Similarly, message characteristics are input and a stochastic process used to generate a message flow for processing. Snapshots of the status of the simulated system may be obtained on four major categories of performance:

manpower utilization, message processing times, workload summary, and an error summary. In addition, a single measure of the system's effectiveness is developed as a composite figure of merit reflecting thoroughness, completeness, responsiveness, and accuracy of the information processing. The MANMOD simulation provides an evaluative vehicle for comparison of alternative system configurations, operational procedures, and personnel characteristics and for diagnostic analyses to identify critical human factors problems for research to form the basis for modifications to the system. Recent improvements in the program have enhanced its capabilities for error simulation and introduced provision for on-line modification of the simulation parameters. The simulation of error processing is predicated on operator identification of the error, based on an error message and recognition of the meaning of the message, and selection of the appropriate corrective procedure. The on-line modification of the simulation parameters permits a user to exercise the simulation and obtain an immediate summary of performance and a detailed description from an off-line printer. A computer-experimenter-subject interactive mode was also introduced to permit use of the program to collect performance data from subjects serving as operators in the system. In this mode, participants perform system tasks as selected by an experimenter, and the obtained data averaged for use in subsequent simulations. Thus, the effectiveness of revised or new procedures may be readily assessed using laboratory obtained performance data.

4. Summary

Human factors in the Army development of battlefield man-computer systems evolved as a joint effort using field evaluations both to identify problems as well as to validate potential solutions resulting from laboratory research. The Tactical Operations System, designed to aid the commander and his staff in decision making, depends on the availability of timely and accurate information for processing and display. The crucial role of the data inputting process to the effectiveness of the system clearly indicates a need to improve human performance (time and errors) of system related tasks. Many human inputting errors are largely undetectable by edit and validation routines, and impact on the system by depriving the user of necessary information or to introduce false information. Furthermore, detectable errors delay the input of information to the extent necessary for correction. Studies under this program have demonstrated the potential benefits of on-line, verified inputting and reference codes designed to facilitate learning and recall. In addition, techniques for error analysis leading to reduced error probabilities, and computer simulation for evaluation and optimization of the system design have been developed.

5. References

[1] Baker, J. D. Acorns in Flower Pots/psychologists in the field. American Psychological Association, JSAS catalog of Selected Documents in Psychology, 1972, 2.

[3] Nystrom, C. O. and G. M. Gividen. Ease of learning alternative TOS message reference codes. U.S. Army Research Institute for the Behavioral and Social Sciences. (In press)

[2] Baker, J. D., D. J. Mace, and J. M. McKendry. The transform operation in TOS: Assessment of the human component. U.S. Army Behavioral Science Research Laboratory, Technical Research Note 212, August 1969. (AD696717)

[4] Strub, M. H. Evaluation of man-computer input techniques for military information systems. U.S. Army Behavior and Systems Research Laboratory, Technical

[5] Strub, M. H. Automated aids to on-line tactical data inputting. U.S. Army Research Institute for the Behavioral and Social Sciences, Technical Paper 262, February 1975. (ADA010350)

[7] Alderman, I. N. and P. A. Gade. Selective feedback in training for on-line tactical data inputting. (In preparation)

Research Note 226, May 1971. (AD730313)

[6] Nawrocki, L. H., M. H. Strub and R. M. Cecil. Error Categorization and analysis in man-computer communication systems. IEEE Transactions on Reliability, Vol. R-22, No. 3, August 1973.

[8] Baker, J. D. Quantitative modeling of human performance in information systems. U.S. Army Behavior and Systems Laboratory, Technical Research Note 232, June 1972. (AD746096)

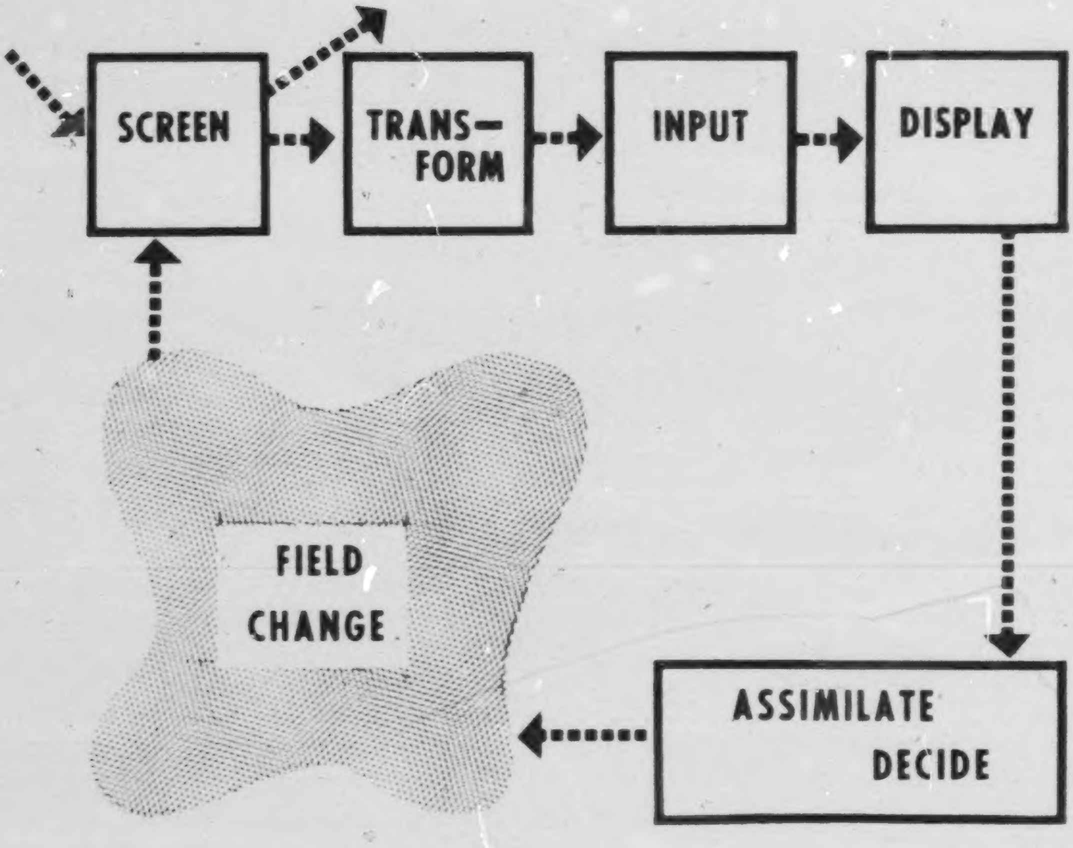


Figure 1.

Schematic representation of operations and information flow in an automated TOS. (Taken from Baker, Mace, and McKendry [2].)

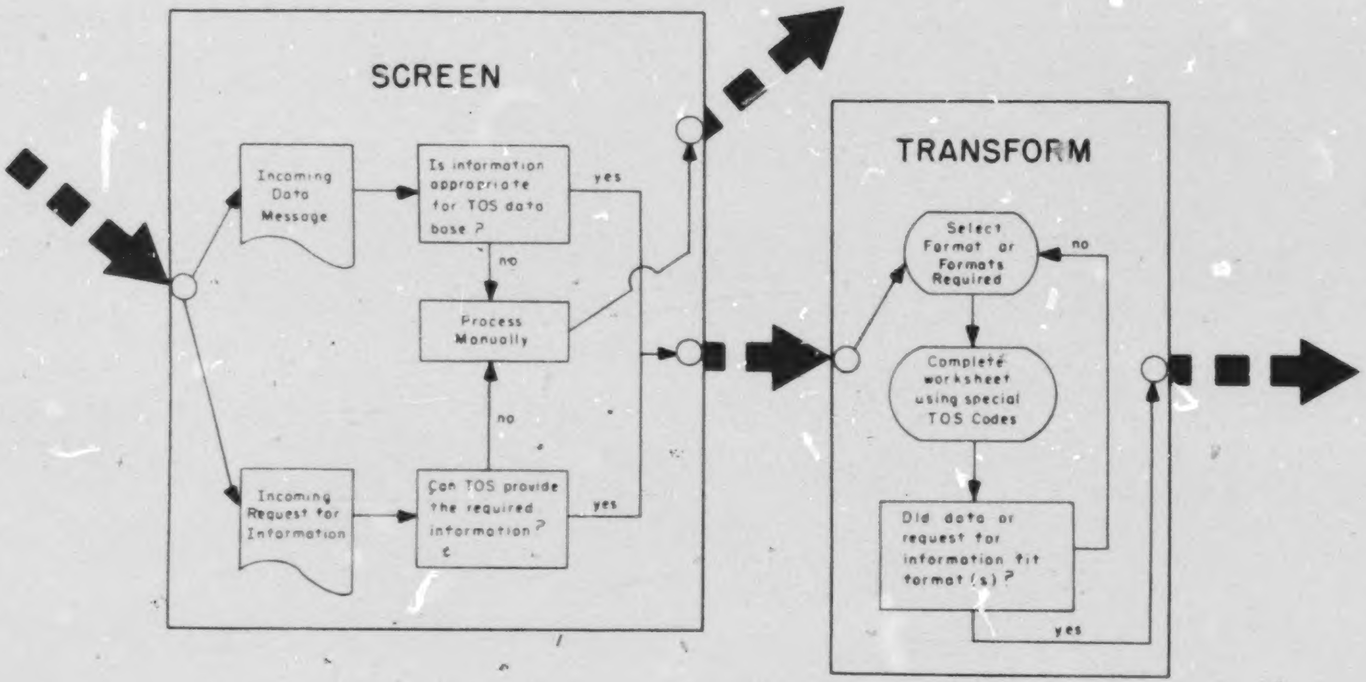

SCREEN
Incoming Data Message
Is information appropriate for TOS data base ?
yes
no
Process Manually
no
Incoming Request for Information
Can TOS provide the required information ?
yes
TRANSFORM
Select Format or Formats Required
no
Complete worksheet using special TOS Codes
Did data or request for information fit format (s) ?
yes

Figure 2.
Schematic representation of the screen and transform operations and data flow in an automated TOS. (Taken from Baker, Mace and McKendry [2].)

"TURNING THE CORNER" ON MIS

A Proposed Program of Data Standards in Post-Secondary Education

Donald R. Arnold, Ph.D.
Director, Office of Management Systems
New Jersey Department of Higher Education

ABSTRACT

There is general agreement that management information systems in post-secondary education have improved substantially during the later decade of the 1960's through to the present.

A major concern is that the full benefit of these advances will not be realized until data processing and management information systems reach a full uniform application of common information units and their expression or representation in standard data formats. Commonality and effective communication can only occur by developing and applying appropriate data standards.

The application and use of common data standards can move management information systems development "around the corner" to finally reach long sought after savings and usefulness in institutional, statewide, and national planning and decision making.

"TURNING THE CORNER" ON MIS

A Proposed Program for Data Standards in Post-Secondary Education

Introduction

In recent years, there has been an enormous expansion in the collection, processing, analysis and exchange of data about the activities, processes and outcomes of post-secondary education. Such information has been essential to the life and operation of the colleges and universities and to statewide and national planning and decision making bodies.

To serve the vital need for improved communication of information among the post-secondary education community, rapid technological advances in computers and management information systems have made it possible to propose and implement increasingly broader integration of data systems and even greater aggregation and reporting of data about the institutions. These advances have achieved some cost reductions and have facilitated important improvements throughout the complete spectrum of data processing systems and services. With only modest exceptions, computers and the technology which they foster have found their way into all of the institutions of post-secondary education.

The Nature of Information Systems

The key to the management information systems concept is the "data base" that is, a massive compilation of hopefully non-redundant, structured and related data stored away that will be instantly accessible to those who have an established need to know, in any combination of relationships, the information residing therein. A management information system can be developed for a single institution. If this were the only objective, specification would require simply that all parts of the institution's data base be capable of relationship to all other parts. From the point of view of statewide and national planning and decision making, the problem becomes somewhat more complex since the specifications for each data element must be agreed upon and the definitions to support these data elements must be uniform throughout the universe. It is only through uniformity that statewide and national accumulation or reporting of information becomes meaningful. To meet the needs of uniformity, there must be an agreed upon minimum set of data elements to insure uniformity of all terms and a data element dictionary to display the logical arrangements of data and the records used.

The lowest level of information systems sophistication is simple record keeping. At this level, the requirement is almost entirely a reflection of local needs, that is, keeping track of students, paying bills, meeting a payroll and other transaction-like activities. All institutions must engage in these kinds of record keeping, but much of it is done manually. At a higher level of sophistication, is the concern for operational and fiscal control. Fiscal control is the primary motive behind line item or objective type budget. It is also the financial control methodology in a majority of the executive offices of state government. Fiscal controls are facilitated by aggregating and re-arranging the kinds of records involved at the lowest level to produce control analyses, indicators, and benchmarks. All institutions are required to maintain fiscal controls, although again, much of this work is usually done without computer support. There is little statewide uniformity of method or format beyond that required to meet the minimum accountabilities under the numerous state statutes or to produce information suitable for annual or bi-annual budgetary review. The information gathered for such a system is often thought of to be inadequate and untimely for meaningful planning and decision making either at the institution or at statewide educational planning levels. The data requires considerable translation usually to programmatic terms and since the definitions of terms and the content is usually not uniform among the constituent institutions, there are frequent errors and much re-checking and re-formatting is required to assure validity and content creditability of the information.

The ultimate objective of a management information system is still one imperfectly realized almost everywhere. This superordinate objective is to furnish information that will permit an executive to judge how resource allocations should be made to meet the overall institutional, statewide or national objectives and to determine how well these objectives are being met. Included are indicators as to how programs and outputs can be measured in cost-effectiveness terms. This requires a level of re-aggregation of data instantaneously into new and flexibly defined relationships and formats. All parts of this system, from individual institution through the statewide and national levels, must translate quickly the financial and program implications of rapid policy or funding shifts. In the final analysis, this level is the most important and is the least well met. It is also the point of which the state-of-the-art in management information systems is least advanced.

The goal of a well convinced management information system is to start with data, build it into records which permit adequate resource and program controls, and, then, to lead decision making regarding planning versus-actual performance in the pursuit of institutional, statewide, and national goals.

A Program of Data Standards

The full benefit of computer systems advances discussed earlier and the imperative for planning and decision making at the institutional, statewide, and national levels, will not be realized until data processing and management information systems reach a uniform application of common information units and their expression or representation in standard data formats. Commonality and effective communication can only occur by developing and applying appropriate data standards. A major concern of institutional decision makers, statewide planners and national policy leaders is that information gathered from the constituent sectors of post-secondary education is frequently highly unreliable as to its content quality and not suitable for comparison either between institutions, among institutional sectors or across differing time periods. All should be concerned that information generated from whatever source is of the highest possible quality and that any differences arising from the data be statistically and numerically significant and not the result of interpretation, definition, or calculation differences. Accordingly, all sectors must exercise a leadership role and undertake a program of data standardization and data definition which is applicable to the entire post-secondary education community.

Data standards are necessary for completion of the Higher Education General Information Survey (HEGIS). After several years of experience with HEGIS, it has become clear that the data are not as usable as might be expected for analysis and planning because the data elements and their definitions are not used or interpreted consistently among the source level institutions. The more typical case is that institutions interpret standards and definitions for local convenience and apply them for their own benefit or self-interest.

The earlier experience with the National Center for Higher Education Management Systems Information Exchange Procedures (IEP) task force re-emphasized the imperative for standard data. For data collected through these IEP procedures to be useful in the decision making and planning processes in post-secondary education, conventions and procedures for gathering and aggregating cost information, must be uniform. Information thus arrived at should be through uniformly defined terms and the reporting and exchange of information should improve communication between the users and providers of information at all levels.

Data Standards are also necessary to link operational data files at the institutional level, but linking as an operational objective is less important as the hierarchy of information builds from the lowest level to the national level. The state-of-the-art in management information systems processing is toward the Generalized Data Base Management System which has as its underlying philosophy the integration of common and uniform data elements among large data files having perhaps dis-similar attributes and resident data. For example, unless a simple attribute like student name, etc., is consistently defined across all files using student name (or its coded equivalent, the student number) the integration of two large files for research or management information reporting purposes is, if not impossible, at least difficult. However, the provision for commonality in the design of a data system is not in itself sufficient. Unless the initiators at the data entry level "fill the blanks" in the consistent and highly specific manner intended, that is, common data elements using common definitions -- only then can common and relatable information be obtained from the system at its output or planning and decision making level.

Early attempts to formulate data standards and definitions have not been successful. Generally, reliance has been placed upon Higher Education General Information Survey reporting system to suffice temporarily while the work of the National Center for Higher Education Management Systems, the National Commission on Financing Post-Secondary Education and the activities among the several states through the State Level Information Base Task Force at NCHEMS, is underway or completed. Management information systems development and proposed data bases have, out of necessity, continued without benefit of a uniform plan. Several large information systems are now installed or in several phases of completion nationally and statewide that prescribe data collection and maintenance routines that "fit the system" rather than those that fit the decision maker or the plan. Data processing systems have evolved and been allowed to become defacto standards without first inquiring as to what use the information thus gathered is to be put and to what extent does the information relate and integrate with other present and proposed systems. The superordinate question is, however, "what data standards and definitions should encompass the entire system of information?"

Colleges and universities are, by definition, collegial entities where decisions seem to be made by consensus and where the concepts of uniqueness and autonomy are paramount. In addition, the concept of application of standard definition presupposes standard behavior or standardization of procedure which has with it the concept of control or accountability. The purpose of this discussion is not to rationalize the environment of post-secondary education, but it is instead to map a course of action out of present practice towards one where standard data elements and standard definitions may work effectively. However, some comments about the collegial approach to developing data processing and information standards is appropriate.

The collegial activity is usually too slow in the deliberation of potential standards and their definition. This is not only a function of the dynamics of the collegial process in general, but frequently the members of the collegium must undertake significant learning processes before they are able to contribute to meaningful debate about the mission of a proposed standard and how the resulting information will be used in the planning and decision making processes.

In addition, members of the collegium tend to function as a committee seeking consensus. Where consensus is not achieved by acclamation, the tendency is to regress to a negotiated position which is usually sub-optimal for the purposes for which the data standard is being considered. Each member of the collegium acts as an individual and brings to the deliberation his or her own views of the environment in which the standard is to operate and frequently argues vehemently for the maintenance of standards which perpetuate the status quo in their own institution or agency.

A Proposed Plan for Action

In this writer's view, the development and maintenance of a plan for data standards must evolve as a policy decision at the highest reasonable level of administrative overview. This policy overview, of course, will be a function of the environment in which the data standard and the resulting information is to evolve and can be either institution wide, statewide or national. It should be understood, at the first order, that the development, maintenance and long time survival of a program of data standards presupposes basic and fundamental behavorial change among persons at the data collection level and pervades the management and administrative processes throughout the entire system. Accordingly, the concept and expected outcome of a program of data standards should be emersed in explicit policy at the highest level.

A proposed structure to facilitate the plan of data standards can be as follows:

1. Seek out available national standards and appraise each standard, it's definition and it's representation as appropriate to the particular need under investigation. Possible sources of pre-defined standards are:
    a) the numerous publications of WICHE/NCHEMS;
    b) the Higher Education General Information Survey (HEGIS) and other National Center for Higher Education Statistics (NCES) publications and procedures;

c) the National Bureau of Standards and Federal Information Processing Standards publications;

d) other -- including publications and procedures pf accrediting agencies, regional and national consortia of educational institutions, bureau of labor statistics, other Federal publications and AAUP, NEA and the like.

A thorough inspection of these documents and publications will reveal that approximately 85% of the information required in the development of a responsible program of data standards which describes the attributes of an institution or a system of institutions in terms of its characteristics, its faculty, its students, its finances, and its facilities, can be uniformly described using available standards.

2. There remains the problem of developing and agreeing upon new standards to meet the undefined 15% which may not have been discovered through the procedures described in the preceding paragraphs. In this instance, it is important to involve the key decision makers and users of information along with appropriate technical council to reach consensus on the unresolved lot of standards. While goal convergence techniques and consensus reaching methodologies are available through such tools as DELPHI, the fact remains that unresolved differences may have to be legislated (or mandated) at the highest level of decision making in the organization.

The activities listed above, that is the inspection of available national standards and the development of new standards for processing, can be an extensive self-analysis about the information gathering and reporting systems of an organization. A typical case is that many items heretofore collected, that may have been defined in the standard manner, that may be maintained in a hard copy file or entered on a machine addressable data base, may be determined, through critical analysis, to be "junk." In these instances, data thus collected which does not meet the criterion for the institutional information program, should be discontinued.

3. When the data standards as a body of technical information have been determined and appropriate definitions applied and their technical requirements for processing in the system of information established, the concept of the specific standard should be codified as institutional or statewide policy. The institutional or statewide data element dictionary, data standards directory and their appropriate taxonomies, should be emersed in institutional policy along with the implementation and/or conversion plan to imbed the program of data standards in the operational routines of the institutions.

4. There remains one additional element which deserves consideration. That is, a follow up institutional policy

and plan that prescribes that no additional data standards may be proposed or used without seeking policy level concurrence having the same purposive decision making practice as was utilized in establishing the standard in the first instance. A program of data standards presupposes compliance with the standard or recognition that the standard exists and knowledge that present processing routines are not now in compliance. Accordingly, it is incumbent upon systems analysts, proposers of data preparation entry routines, and decision makers to first inquire whether data elements exist in the standard form and definition before proposing new processing routines which manipulate data or cause data to be collected and entered at the lowest level.

Summary

A concept of developing of program of data standards and their definitions: 1) that is imbedded first in institutional or statewide policy; 2) that seeks out and agrees upon available national standards; 3) that then sets the standards as operational policy for the institution; 4) that is committed operationally, and 5) that has appropriate follow-up and audit procedures, will move management information systems and computer technology into the "golden age" of usefulness in the planning and decision making processes in post-secondary education. It should be understood, however, that a program of data standards fundamentally causes or effects the behavior of the members of the organization as well as the structure of information flow and communications. A final caution would be that inasmuch as people are effected and behavior change is likely prescribed, the analyst, the decision maker, and the policy maker must understand that a comprehensive program of data standards is expensive to implement and may require extended timeframe for implementation.

SELECTED BIBLIOGRAPHY

Dahnke, Harold L.; Jones, Dennis P.; Mason, Thomas R.; and Romney, Leonard C. Higher Education Facilities Planning and Management Manuals. Technical Reports 17-1 through 17-7. Boulder, Colorado: Western Interstate Commission for Higher Education, May 1971.

Goddard, Suzette; Martin, James S.; and Romney, Leonard C. Data Element Dictionary: Course, Facilities, Finance, Staff, Student (Second Edition). Technical Report 51. Boulder, Colorado: Western Interstate Commission for Higher Education, November 1973.

Gulko, Warren W. Program Classification Structure. Technical Report 27. Boulder, Colorado: Western Interstate Commission for Higher Education, January 1972.

Johnson, Richard S. and Huff, Robert A. NCHEMS Information Exchange Procedures (Preliminary Edition). Boulder, Colorado: Western Interstate Commission for Higher Education, May 1974.

Management of Data Elements in Information Processing, Proceedings of the First National Symposium. U.S. Department of Commerce, National Bureau of Standards, Washington, D.C. 1974.

Manning, Charles W. and Romney, Leonard C. Faculty Activity Analysis: Procedures Manual. Technical Report 44. Boulder, Colorado: Western Interstate Commission for Higher Education, 1973.

National Task Force on Student Aid Problems, Final Report, 1974.

Minter, John W. A Manual for Manpower Accounting in Higher Education (Preliminary Edition). U.S. National Center for Educational Statistics. Washington, D.C.: Government Printing Office, 1972.

Romney, Leonard C. Higher Education Facilities Inventory and Classification Manual. Technical Report 36. Boulder, Colorado: Western Interstate Commission for Higher Education, December 1972.

U.S. Department of Commerce, National Bureau of Standards, Counties and County Equivalents of the States of the United States. FIPS Pub 6-2, June 1970.

U.S. National Center for Educational Statistics. Encyclopedia of Education. Washington, D.C.: Government Printing Office, 1971.

ASCII - The Data Alphabet That Will Endure

Robert W. Bemer
Honeywell Information Systems, Inc.
Phoenix, Arizona, US

A standard data alphabet is indispensable to understanding communication and reading data in machine-encoded form (not spoken, not written or printed). ASCII (the ISO Code) has, by design, capabilities for expansion and extension not inherent in any other code. The many billions of dollars worth of ASCII-based communication and computation equipment is the best prepared for the coming fields of networking, electronic funds transfer, text processing and photocomposition, the automated office, etc.

The status and prospects of this healthy 12-year-old are explored.

Keywords: Alphabet; ASCII; character; code; ISO; symbol

1. ASCII is a True Alphabet

Because ASCII [1] and its international identical twin, the ISO Code [2], are actually called "coded character sets" in the formal standards, I must begin by explaining why I use the term "alphabet" instead. One reason is that another international identical twin is called the C.C.I.T.T. Working Alphabet No. 5. For the others, some definitions are necessary. Even though I dislike Webster's Third International (sic) Dictionary intensely, here are some of the things it says that "alphabet" means:

1a. Any particular set of letters with which one or more languages are written, especially such a set of letters arranged in customary order.

1b. Any set of characters with which one or more languages are written, whether these characters are letters (sense 1a), signs of a syllabary, or other basic units of writing.

1f. The alphabetic system of writing, as distinguished from syllabic, ideographic, and other systems.

1h. Any system of signs or signals, visual, auditory, or tactile, that serve as equivalents for the usual written letters of the alphabet.

1i. A particular set of names used to designate the various letters in the alphabet (the pronouncing alphabet used in civil aviation).

1j. In cryptology, a set of one-to-one equivalences between a sequence of plaintext letters and the sequence of their cipher substitutes.

Are not ASCII and the ISO Code actually alphabets in every such sense?

In sense **1a**, it contains letters, and they are arranged in the customary order (but not collating sequence, because of the dual case representations of the letters) by the numerical order of their bit encodings. More than one language can be written with it; as the international code, the most prevalent languages using the Roman alphabet can be written. Note particularly, in the official reference version, that provision is made for the extra Scandinavian letters, located in the proper position (although the usage is not the same in Denmark, Norway, and Sweden).

In sense **1b**, it contains other units of writing. Punctuation is there, as are underscore and other common symbols. Diacritical elements exist for forming compound and accented letters, thus bringing more Roman-based languages within its capability.

In sense **1f**, it certainly does not have syllabic or ideographic characteristics. So it is not excluded from being called an alphabet for these reasons.

In sense **1h**, the encoded representations are the equivalents. In fact, this is the definition of most interest to us. Note that after 90 years of encoding (starting with punch cards for the 1890 Census), Webster's Dictionary fails to give specific status to this manifestation.

Senses **1i** and **1j** may not appear to be pertinent now, but they are there for a reason, and we shall return to them.

2. Why Alphabets?

Alphabets comprise a class of methods to record knowledge for transmittal to others. To transfer knowledge we must transfer information; to transfer information we must transfer data.

There are, of course, other methods using basic elements at higher levels of complexity -- such as syllables, ideographs, etc. However, the primitive ASCII (ISO Code) is *the* worldwide standard for exchange of data and information. It continues to be the standard because the primitives are representable in alternate but related ways.

It is interesting to speculate what might have happened if our forebears had developed the phonograph or tape recorder *before* writing with alphabets. Would we be as deeply into databanks as we are now? Would standards have been developed for speech sounds in analog form, using the computer to discriminate and remove differences in people's voices? Would Confucius have said that one analog picture was worth a thousand digital words?

One suspects it might be difficult to search such a databank, however. At least it appears that we do not have the methods yet. For example, in reply to the Senate Committee asking for information from the Nixon tapes, the White House says that they have not been classified as to content, and it would take listening to them in their entirety. In other words, a linear realtime search -- and we know how inefficient that can get as databases grow larger.

3. Why ASCII Survives and Grows

My personal history or view of the development of the ISO Code and ASCII [3] tells of millions of dollars of careful international effort and planning spent in its creation. But that is minor, amortized by many billions now invested in communications and computer equipment that operate via ASCII. It also tells of the IBM code called EBCDIC, a result (according to Fred Brooks, one of the chief designers of the 360) of forced announcement before ASCII peripheral equipment could be completed. Although the 360 was *said* to have ASCII capability, it was never realized in the software.

If computers, in substantial portion, do not operate in ASCII as native mode, then why will EBCDIC not be the survivor? Many people, both in and out of government, have blithely assumed that EBCDIC will -- and continue to invest money in software and operations based upon EBCDIC.[4] They are going to be very surprised, because *IBM knows* that EBCDIC will eventually be subordinate! [5]

Why? Because EBCDIC is not, like ASCII, the result of meticulous design.

> "... it would appear that no single 'computer code' can be completely adequate, and that insistence on a single code for all purposes would be counterproductive. Rather, the Federal Government should maximize the benefits to be accrued from taking advantage of our growing technological ability to live in a multi-code world ..."[5]

In ASCII, the controls are all located in the leftmost two columns. It is compact, extensible, expandable, and even subsettable. It can grow easily into an 8-bit code (expandability), or into 9-bits, 10-, or anything. At any level of byte size, it can be extended to encompass alternate sets of characters; keeping the same control columns, various pages can be substituted for the other columns. The methods for expansion and extension are also standardized.[6,7] Sets having sufficient utility may be registered for international usage, via the French standards body AFNOR, which holds the secretariat for international code standardization within ISO TC97. The vehicle for doing this is the ESCape character.[8] Various pages are registered with unique ESCape sequences.[9,10]

4. Code Extension

In the extension procedures, the existing 7-bit ASCII is divided into control and graphic portions. The first two columns of code -- the controls -- comprise the C0 set; the other six columns -- the graphics -- comprise the G0 set. The extended set first removed from basic ASCII is similarly divided into the C1 and G1 sets. Obviously such sets could be adjoined in the 8-bit form, and the USSR [11] and Japanese [12] standards are excellent examples of so doing.

4.1 Extended Control Sets

C1 sets can be, and have been, designed for many purposes. The one furthest progressed to agreement is that for softcopy controls, for CRT display screens [13,14].

> "... The major difficulties at present are in using the established 'control' characters with devices that had not been invented at the time the code was, or in extending the 'graphic' symbol set to meet new application requirements ..."[5]

Work is continuing in both ECMA (the European Computer Manufacturers Association) and ANSI X3 to get agreement sufficient for final registration. The original work of X3L2 was for softcopy controls to be in an *expanded* set (8-bit code), but that is presumptious.

				b_7	0	0	0	0	1	1	1	1
				b_6	0	0	1	1	0	0	1	1
				b_5	0	1	0	1	0	1	0	1
b_4	b_3	b_2	b_1		0	1	2	3	4	5	6	7
0	0	0	0	0			SP	0		P		
0	0	0	1	1				1	A	Q		
0	0	1	0	2			"	2	B	R		
0	0	1	1	3				3	C	S		
0	1	0	0	4			$	4	D	T		
0	1	0	1	5				5	E	U		
0	1	1	0	6				6	F	V		
0	1	1	1	7				7	G	W		
1	0	0	0	8			(	8	H	X		
1	0	0	1	9			)	9	I	Y		
1	0	1	0	10			*		J	Z		
1	0	1	1	11			+	;	K			
1	1	0	0	12			,	<	L			
1	1	0	1	13			-	=	M			
1	1	1	0	14			.	>	N			
1	1	1	1	15			/		O			

Figure 1. COBOL Character Set

				b_7	0	0	0	0	1	1	1	1
				b_6	0	0	1	1	0	0	1	1
				b_5	0	1	0	1	0	1	0	1
b_4	b_3	b_2	b_1		0	1	2	3	4	5	6	7
0	0	0	0	0			SP	0		P		
0	0	0	1	1				1	A	Q		
0	0	1	0	2				2	B	R		
0	0	1	1	3				3	C	S		
0	1	0	0	4			$	4	D	T		
0	1	0	1	5				5	E	U		
0	1	1	0	6				6	F	V		
0	1	1	1	7			'	7	G	W		
1	0	0	0	8			(	8	H	X		
1	0	0	1	9			)	9	I	Y		
1	0	1	0	10			*	:	J	Z		
1	0	1	1	11			+		K			
1	1	0	0	12			,		L			
1	1	0	1	13			-	=	M			
1	1	1	0	14			.		N			
1	1	1	1	15			/		O			

Figure 2. Fortran Character Set

ISO Technical Committee 46 (Documentation), in its Subcommittee 4 (Automated Documentation) has a working group on bibliographic codes. Its first candidate for registration as a C1 control set is a set for bibliographic controls [15] to be embedded in text to delimit certain special data. This C1 set contains four classes of characters – annotation controls, filing controls, reference controls, and subject designators. Major credit should be given here to Dr. Ernst Kohl of the Bavarian State Library in Munich.

Although little work has been done, other C1 control sets are envisioned for typographic control – to vary the font, weight, slope, size, and spacing, etc. of the graphic characters. Other sets could be envisioned for fields such as process control, animation and other graphics applications, sewing machines, etc. Do you think the last one far-fetched? Singer has already announced a machine with a microcomputer, and is there any reason to think that future models won't use ASCII characters, in a hand calculator type of display, to give instructions and options available?

4.2 Extended Graphic Sets

G1 sets are further along. ISO TC46/4/1 has tabled Draft International Standards for Latin (DIS-5426), Greek (DIS-5427), and Cyrillic (DIS-5428). We may presume the latter is in harmony with [11]. Under study are sets for mathematical characters and the African languages. Proposals have been solicited for such languages as Arabic, Kata Kana, Kanji, etc. It is permissible for a G1 set to be a partial replication of the basic G0 set of ASCII; indeed, many are very similar, with the lower case being replaced by the new alphabet.

Although we have seen the ISO assignments for natural languages to be in the jurisdiction of TC46, TC97 (Computers and Information Processing) has retained authority to make assignments for programming languages. The work has been concentrated in G1 sets for COBOL (figure 1), Fortran (figure 2), Basic (figure 3), and PL/I (figure 4). A table for APL is being constructed. ALGOL presents different problems.[16]

				b_7	0	0	0	0	1	1	1	1
				b_6	0	0	1	1	0	0	1	1
				b_5	0	1	0	1	0	1	0	1
					0	1	2	3	4	5	6	7
b_4	b_3	b_2	b_1									
0	0	0	0	0			SP	0		P		
0	0	0	1	1			!	1	A	Q		
0	0	1	0	2			"	2	B	R		
0	0	1	1	3			#	3	C	S		
0	1	0	0	4			$	4	D	T		
0	1	0	1	5			%	5	E	U		
0	1	1	0	6			&	6	F	V		
0	1	1	1	7			'	7	G	W		
1	0	0	0	8			(	8	H	X		
1	0	0	1	9			)	9	I	Y		
1	0	1	0	10			*	:	J	Z		
1	0	1	1	11			+	;	K			
1	1	0	0	12			,	<	L			
1	1	0	1	13			–	=	M			
1	1	1	0	14			.	>	N	^		
1	1	1	1	15			/	?	O	_		

Figure 3. BASIC Character Set

				b_7	0	0	0	0	1	1	1	1
				b_6	0	0	1	1	0	0	1	1
				b_5	0	1	0	1	0	1	0	1
					0	1	2	3	4	5	6	7
b_4	b_3	b_2	b_1									
0	0	0	0	0			ƀ	0		P		
0	0	0	1	1			|	1	A	Q		
0	0	1	0	2				2	B	R		
0	0	1	1	3				3	C	S		
0	1	0	0	4			$	4	D	T		
0	1	0	1	5				5	E	U		
0	1	1	0	6			&	6	F	V		
0	1	1	1	7			'	7	G	W		
1	0	0	0	8			(	8	H	X		
1	0	0	1	9			)	9	I	Y		
1	0	1	0	10			*	:	J	Z		
1	0	1	1	11			+	;	K			
1	1	0	0	12			,	<	L			
1	1	0	1	13			–	=	M			
1	1	1	0	14			.	>	N	¬		
1	1	1	1	15			/		O	_		

Figure 4. PL/I Character Set

One may be tempted to think of these not as G1 sets but rather as subsets of the G0 set, standard ASCII. But note that they are incompatible in minor ways, particularly for PL/I. which was the cause of considerable difficulty in stabilizing ASCII. So perhaps the G1 status is an easy solution.

4.3 The Registry Method

A responsible standardizing body with a specific proposal makes application to AFNOR, acting as agent for ISO TC97/SC2. Applications may be for graphic sets (G1, etc.), C0 or C1 control sets, a single control character, or a code requiring special interpretation. The approval procedure is defined in [10]. A unique ESCape sequence is assigned. It is here that definitions *1i* and *1j* for alphabet become applicable. The ESCape sequence, as adjoined to any following character before termination, becomes a name for the alternate characters and alphabet, in one-to-one equivalence. Thus all of the world's symbol and alphabets may be represented uniquely for interchange.

5. How IBM Can and Will Use ASCII

EBCDIC is a sparsely settled code that utilizes the 8-bit capability of 256 characters ineffectively. The collating sequence(s) are not easily derivable from the numerical values of the coded representations. The controls are intermingled with the other characters, so that it can not be extended by paging, as ASCII can. It has only one redeeming virtue -- one-to-one correspondence with ASCII via a common character set as represented in punch cards! (See figures 5, 6)

> "The interesting observation is that if two character codes each have the same symbol set, and if each meet the requirement of no symbol ambiguities for the same bit pattern (no duals), then automatic context-free translation between the two character codes is a trivial task ... The operating cost of translation (between two such character codes) concurrently with preparing or accepting an interchange message is trivial in today's systems and will be more so in tomorrow's LSI machines."[5]

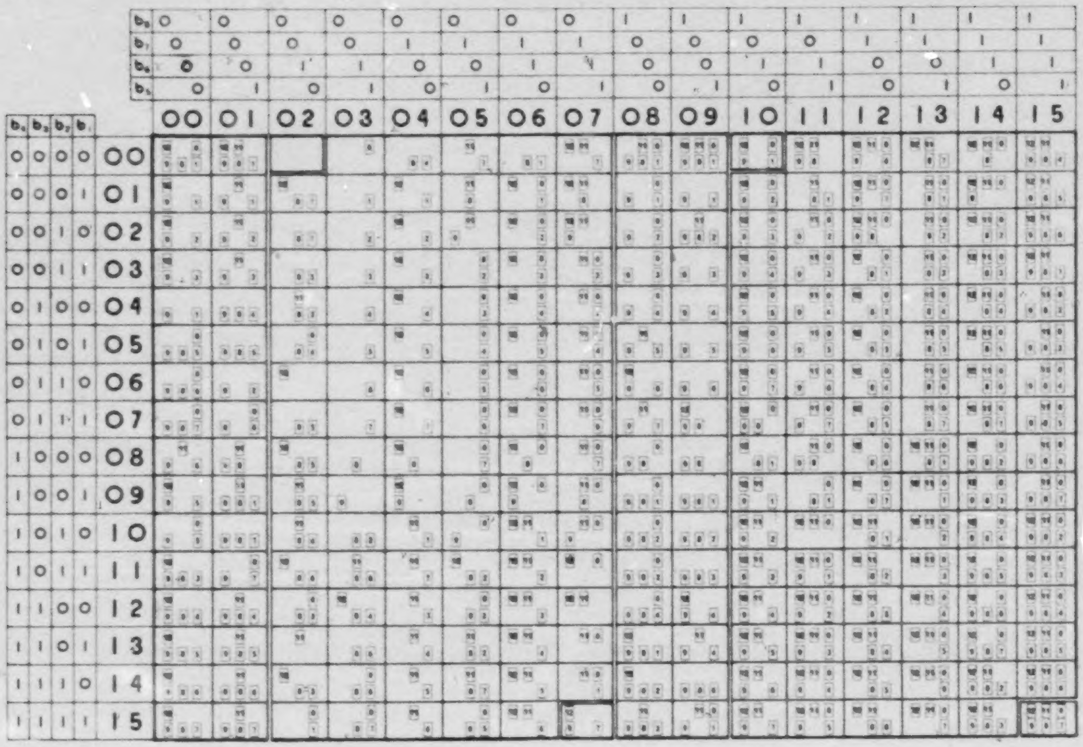


Figure 5. Hole Patterns Assigned By Code

	12	11	0		12 0	12 11	11 0	12 11 0
	02/06	02/13	03/00	02/00	07/11	07/12	07/13	11/10
1	04/01	04/10	02/15	03/01	06/01	06/10	07/14	13/09
2	04/02	04/11	05/03	03/02	06/02	06/11	07/03	13/10
3	04/03	04/12	05/04	03/03	06/03	06/12	07/04	13/11
4	04/04	04/13	05/05	03/04	06/04	06/13	07/05	13/12
5	04/05	04/14	05/06	03/05	06/05	06/14	07/06	13/13
6	04/06	04/15	05/07	03/06	06/06	06/15	07/07	13/14
7	04/07	05/00	05/08	03/07	06/07	07/00	07/08	13/15
8	04/08	05/01	05/09	03/08	06/08	07/01	07/09	14/00
9	04/09	05/02	05/10	03/09	06/09	07/02	07/10	14/01
8 2	05/11	05/13	05/12	03/10	12/04	12/11	13/02	14/02
8 3	02/14	02/04	02/12	02/03	12/05	12/12	13/03	14/03
8 4	03/12	02/10	02/05	04/00	12/06	12/13	13/04	14/04
8 5	02/08	02/09	05/15	02/07	12/07	12/14	13/05	14/05
8 6	02/11	03/11	03/14	03/13	12/08	12/15	13/06	14/06
8 7	02/01	05/14	03/15	02/02	12/09	13/00	13/07	14/07

12	11	0		12 0	12 11	11 0	12 11 0	
10/06	11/01	11/09	06/00	12/03	12/10	13/01	13/08	8 1
00/01	01/01	08/01	09/01	10/00	10/09	09/15	11/11	9 1
00/02	01/02	08/02	01/06	10/01	10/10	11/02	11/12	9 2
00/03	01/03	08/03	09/03	10/02	10/11	11/03	11/13	9 3
09/12	09/13	08/04	09/04	10/03	10/12	11/04	11/14	9 4
00/09	08/05	00/10	09/05	10/04	10/13	11/05	11/15	9 5
08/06	00/06	01/07	09/06	10/05	10/14	11/06	12/00	9 6
07/15	08/07	01/11	00/04	10/06	10/15	11/07	12/01	9 7
09/07	01/08	08/08	09/08	10/07	11/00	11/08	12/02	9 8
08/13	01/09	08/09	09/09	00/00	01/00	08/00	09/00	9 8 1
08/14	09/02	08/10	09/10	14/06	14/14	15/04	15/10	9 8 2
00/11	08/15	08/11	09/11	14/09	14/15	15/05	15/11	9 8 3
00/12	01/12	08/12	01/04	14/10	15/00	15/06	15/12	9 8 4
00/13	01/13	00/05	01/05	14/11	15/01	15/07	15/13	9 8 5
00/14	01/14	00/06	09/14	14/12	15/02	15/08	15/14	9 8 6
00/15	01/15	00/07	01/10	14/13	15/03	15/09	15/15	9 8 7

Figure 6. Code Assigned To Hole Patterns

Anyone with a $15 hand calculator realizes how cheap a microelectronic chip must be, for his calculator does a more complex job than the job of converting back and forth between the ASCII and EBCDIC encodings for the same character.

So let us postulate a very inexpensive chip inside IBM computers. It converts from EBCDIC to ASCII or from ASCII to EBCDIC without any delay as input-output or other operations are executed. Two questions must be asked:

- Is the data EBCDIC or ASCII?
- Does the program expect EBCDIC data or ASCII data?

Imagine subscripting the comparison instruction by EBCDIC or ASCII tags. I have a master file in EBCDIC, against which I run an update tape in ASCII. My program says "Compare the keys on an ASCII basis". The CPU, noting that the key of the master file is in EBCDIC, routes it through the chip before attempting to compare it to the key from the update record. My program can also give instruction to convert the entire updated file to ASCII automatically as it is being stored.

Thus the feasible technique. How about some signs that it will be so?

- Note IBM's commitment to word processing and photocomposition. According to B. O. Evans:

> "There are similar requirements for the ability to use differing codes in differing contexts to represent different graphic needs. In particular, the development of 'end user' devices, such as photocomposers and interactive displays, requires greatly expanded symbol sets to be developed for some applications -- more symbols than can be contained directly in a 7- or 8-bit code. Thus, specific codes (of however many bits are required to represent a symbol) might well be developed for data interchange in certain application areas utilizing such devices."[5]

- Note IBM's 6250 cpi magnetic tape, which departs from cross-tape parity checking and recognizability of code without programmed knowledge.

- Note IBM's firm and continuing insistence that data control procedures be bit-transparent and not byte-oriented -- particularly not 8-bit-byte-oriented.

> "We expect to see machine architectures having the flexibility to adapt efficiently to as-yet-undefined code structures without disruption of existing applications at any time even after the system was installed."[5]

- Note SDLC and SNA. When you are going to be communications-oriented, and even run a satellite system, why object to using the ASCII code that *all* communications is based on?

6. Conclusions

ASCII was well-designed, and is flexible to adapt to usage that may take any turn of development. There is no reason, with presently available technology, to use any other encoded alphabet. As a single standard, it enables private data to become public whenever that is desirable (i.e., privacy may be protected or maintained in ways other than unintelligibility). It is the alphabet of all communications networks, of all minicomputers, and of some larger computers such as the NCR Century series.

After twelve years it is still healthy, and when IBM puts the seal of approval on it (viz. virtual memory and APL) it will be the undisputed universal interchange medium and linguist.

7. References

1. ANSI X3.4 - 1974, "American national standard code for information interchange"
2. ISO 646 - 1973, "6 and 7 bit coded character sets for information processing interchange"
3. R.W.Bemer, "A view of the history of the ISO code", Honeywell Comput. J., **6**, No. 4, 274-286 (1972)
4. Federal Information Processing Standard Publication **7**, "Implementation of the code for information interchange and related media standards"
5. B.O.Evans, letter of 1973 June 28 to R.R.Johnson, Chairman Federal Information Processing Standards Task Group 12.
6. ANSI X3.41 - 1974, "Code extension techniques for use with the 7-bit coded character set of (ANS) code for information interchange"
7. ISO 2022 - 1973, Code extension procedures for ISO 7 bit code"
8. R.W.Bemer, "Escape - a proposal for character code compatibility", Commun. ACM **3**, No. 2, 71-72 (1960 Feb)
9. R.W.Bemer, Working Paper, "Registry - A new feature for ISO recommendations", 1970 May 28
10. ISO 2375 - 1974, "Procedures for the registration of escape sequences"
11. USSR State Standard, GOST 13052-67, "Computers and data transmission equipment, alpha-numeric codes", 1967 July 10
12. Japanese Industrial Standard Code for Information Interchange, "JISCII", C6220-1969
13. T.O.Holtey and E.H.Clamons, "Soft-copy controls", Honeywell Comput. J. **7**, No. 4, 267-269 (1973)
14. ANSI X3L2/1499, "Proposed American national standard - control characters for an 8-bit code", 1975 Oct 10
15. ISO/TC46/4/1 N54, "The set of specific bibliographic control characters", 1975-08-12
16. ISO Third Draft Proposal 1672, "Hardware representation of ALGOL basic symbols in the ISO 6 and 7 bit coded character sets"

Techniques in Developing
Standard Procedures for Data Editing

George W. Covill

Automation Industries, Inc.
Vitro Laboratories Division
14000 Georgia Avenue
Silver Spring, Maryland 20910

PLAYSCRIPT Procedures are a method of presenting information simply to employees who are engaged in the pre- and post-processing activities of a data processing system. It can also be applied to explain and detail other administrative activities in support of data processing.

Key words: Documentation; PLAYSCRIPT; procedure; standard.

Machines can process tremendous quantities of data in minimal amounts of time, but the quality of this data is still subject to the original axiom: garbage in, garbage out! Advances in the art of machine editing have alleviated one part of the problem; not solved it. The necessity of some human reviewing computer input/output for the purpose of validating accuracy is still with us. Humans are like machines in one aspect: if they are to do something they must be told what it is they are to do. Verbal instructions are often sufficient, but in most cases, an organized and simple written exposition is best. This paper presents an approach used successfully in communicating a standard procedure for the editing of scientific data to non-scientific oriented editing personnel.

In 1970 Vitro Laboratories assumed the responsibility of managing a large data bank of chemical research data on tumor chemotherapy. While the company staff had accumulated an impressive background in the management of technical data banks since 1950, this particular application was to be a new experience for the assigned task personnel. The data came from a multiplicity of sources around the world, with wide variation in data volume, accuracy, and skill of the data originator. The new application required current processing of a high volume of discrete items of clinical data which had to meet acute standards of accuracy in validation and input preparation. The high standard of data inputs had to be achieved at all times, even when turnover occurred, due either to promotion or termination.

The computer programs embodied very sophisticated editing routines at each step of the file updating and reporting process, and the system was batch-processing oriented on a weekly or bi-weekly schedule. Therefore, although errors in pre-processing editing would normally be rejected somewhere along the line, it would be at least two weeks, and sometimes four, before error correction could be accomplished due to some types of errors requiring reference back to the originator for correction. In turn, this delay would further retard the recording and reporting process, creating timing difficulties in an on-going research project.

The solution, of course, was to investigate, standardize, and document the pre-processing procedures. In accomplishing the investigation process, the analysts were aided by interviews with employees and supervisors already performing the tasks, advice of the programmers developing the programs, and not least of all, consultation with, and advice of, the customer. Not far into this process, a determination was made that a more effective method of presenting the editing information to the ultimate user was required. In the normal course of events, such details as were required in pre-processing would be consolidated into a book which would terrify the intended users by its imposing size, and confuse them by the difficulty of locating that part, or those parts, which applied specifically to their particular task. What was needed was a pamphlet covering each specific task in the pre-processing sequence which could be handed to a new employee, and by its small size and lack of complexity indicate that the job was not beyond normal human ken.

The method selected was called "Playscript" procedure. Its progenitor, Leslie Matthies, has authored a book called, oddly enough "The PLAYSCRIPT Procedure".[1] Anyone who may plan to use the PLAYSCRIPT procedure would be well advised to secure a copy of this book, read it carefully, and apply the principles enunciated in it faithfully.

What is a PLAYSCRIPT? Well, if you have ever read a book of plays you already have a grasp of the basic idea. At this point take a look at the attached procedure 02.13-2, Editing of Control Packs in Vivo. You can see that two different forms are used in its preparation, and that the areas where certain standard types of information is to appear are formally delineated. All of this formality is really unnecessary. Equally effective procedures can be, and are being, prepared on plain white paper. All that is really essential is standardization of the typing format so that there is a clear relationship in numbering and titling established between one procedure and all the others. Furthermore, practically all of the information in paragraphs I, II, and III on the first page can be dispensed with. Paragraph II. GENERAL helps some, as it really is a stage setting for the instructions to follow, and serves as a catchall where information not easily placed into PLAYSCRIPT format can be published.

Paragraph IV. PROCEDURE, contains the real heart of a PLAYSCRIPT, a step-by-step presentation of the action necessary to complete a job. There are four basic parts to each PLAYSCRIPT action: an actor's name, a sequence number, an action word, and a complete action sentence. Looking at page 2 of the procedure you will see how each is employed. An actor's name (appearing under Responsibility) is not an actual person's name, but really the job title of the person doing the job being described; the sequence number identifies each step to be taken in accomplishing the total job; the action word is an active verb in the present tense which characterizes and precedes the action to be taken in each particular step; and the complete action sentence specifies what the actor does step-by-step. If the playout of the action is transferred to another actor, his title appears and the narration continues. These four items are the basic building blocks of a PLAYSCRIPT procedure, and can be used to describe in detail the most complicated tasks in editing data, or in accomplishing administrative tasks. A glance at a completed procedure indicates its basic simplicity. However, this simplicity is very difficult to achieve in practice. Let me say that our simplicity improved with practice, without necessarily being completely achieved.

The preparation of effective procedures requires standardization of format and content before starting; otherwise, some initial effort will have to be repeated in order to achieve the expected results. Decisions must be made as to the general appearance of a procedure, i. e., should it be on special forms (not really necessary, nor even desirable), whose signature(s) shall indicate approval, where will the signatures appear, what will be the standard of quality of art work (if any), what are the requirements for coordination of procedures prior to publication, and so on. These decisions do not effect content, but they do affect the attitude with which the published procedure will be received, and whether supervisors will just file and forget them, or actually use them in accomplishing the job. Let's say right here, that if the supervisor is not actively involved in some way in the development of a procedure affecting his operation, the procedure has no future. Conversely, the more active the supervisor, the better the procedure, and the more accurate and useful it is on the job.

---

[1] Matthies, Leslie H.; The Playscript Procedure: A New Tool of Administration; New York: Office Publications, 1961.

Once it is decided that formal procedures are to be utilized, their production must have an organized objective. First of all, a list of the general subjects to be covered must be outlined. Each procedure to be written must then be given a name and a number, the number so that there will be an organized scheme for filing, and a title to identify content. The number of the attached procedure is 02.13-2, which indicates that it is the 13th procedure in the series describing the editing of control packs, and that this is the second published revision. Also it helps to begin with an organized cast of actors, permitting the standardizing of the titles to appear in the procedures, and also providing a clue to the types of work to be covered by the procedures. The actor's name is significant in that people like to see their particular job title in writing, and knowing it before you approach the worker helps in confirming that you have done your homework. Also, a collection of blank and completed forms and reports which are to be the basis of the work to be described should be in hand and organized. A study of these will help to make the nature of the task much clearer. Another helpful file is any previous documentation of the background and use of the forms and reports and their contents. Any information which the analyst can acquire prior to approaching the task of actually describing the job at the work place will pay off in a more enthusiastic reception by supervisor and worker, a better understanding of what the analyst is verbally told about the job, and the more useful the resultant product.

The first major hurdle to be taken by the analyst is that of gaining the confidence of the supervisor. Depend upon it, the initial reaction will be that "some hotshot is going to come down and tell us how to do the job." It is difficult to get across that the analyst is really there to listen, and learn, not to pontificate. A preliminary visit to all supervisors explaining the objectives of the program, and their part in it, can help. Another big help can be that the supervisor's supervisor is behind the program. In any case, the analyst must be careful in what he does at the workplace, and even more careful in what he says, for even the most innocent remark can be interpreted as a slur when the will is present.

What the analyst's main objective must be is to actually get to the workplace and participate with the employee in the process of doing the job. Only then will he be able to get the full details. Remember, most employees find difficulty in articulating about their job functions. Others are prone to slight what they consider minor repetitive tasks. Ask questions, repeat them where necessary, exhibit a healthy curiosity, and be appreciative of the importance of the individual's task. Remember, the analyst is a guest in another person's environment, and should modify his personal conduct accordingly.

Above all, be careful about making notes. There is a basic resentment about having one's remarks written down and possibly quoted later to cause embarrassment. If notes are taken, let the subject read them, or better yet, read them to him, before leaving the area. This will not only relieve possible embarrassment, but will serve to clarify any areas of misunderstanding.

Then prepare the preliminary procedure. Write it in just the way the job appears to be done. Don't try to improve on the worker's performance. If possible areas of improvement are discerned, write them also, but not in the preliminary procedure. Stick to the facts obtained in the interview. In developing the procedure, stick to the positive approach. There isn't any need to present the wrong way to do the job, tell why it is wrong, and then give the proposed way. What we are aiming to do is teach the proposed improvement, the errors will occur of themselves without instruction.

Now go over the preliminary procedure with the supervisor. If there should be areas of disagreement, try to have the actual process verified at the work station, not thru discussion. Modify the procedure if necessary, and secure agreement in the modifications. Now is the time to bring up the subject of possible improvements in the process, but be careful. Unless some rapport exists between analyst and supervisor, these well meant suggestions will be regarded as criticisms. Remember, the supervisor is responsible for getting the job done. No matter how pertinent the improvements, if he isn't interested they will never be implemented. Wait for a better day.

Suggest to the supervisor that he and the actor (employee) read and discuss the preliminary procedure together, and leave the procedure with him. Check back later to discuss any possible changes. Particularly verify the accuracy of any examples of data in the text, and of the exhibits.

Now write the final copy. Do not change what has been agreed upon without going back for reverification. Remember, the supervisor and worker must have confidence in you; anything which can be viewed as a breach of trust can be fatal to this relationship.

Prior to publication have the final copy again reviewed by the supervisor and by the worker, if possible. When everyone is agreed, then the procedure is ready to be sent to the customer for final review and approval. Upon return, the procedure is officially ready for implementation. If at all possible, the analyst should see that the supervisor concerned receives a personal copy prior to receipt thru normal distribution. These little touches of personal concern are appreciated.

We found that there were certain advantages to having standard procedures. First of all, they did provide a method of normalizing the editing procedure. When errors of omission in editing were discovered, and were described in a procedure, reference could be made to the applicable publication, with a caution to the editor to follow it. Secondly, the actual quantity and quality of work required to process the incoming material was established. If anyone wanted to, and at times they did, the procedure could be stepped thru and timed. Using this unit of measure the actual impact of changes in work volume could be easily established. Thirdly, continuation of effort was maintained thru changes in personnel due to illness, vacation, termination, etc., along with quality standards. Fourthly, a channel of communication was opened between customers, contractor, and worker by which desired changes in job procedures were accurately documented and communicated. Not least of all, the production of procedures was an on-going project covering two and a half years, during which each of the original procedures was updated more than once, so that the experience and suggestions of a number of people actually doing the job become incorporated into subsequent versions.

During the period of the contract the number of required procedures rose from 36 to 92. This was because the success of the initial effort caused expansion into areas not originally considered. Standard procedures were prepared for almost all of the administrative activities involved in the contract, and particularly in those areas where difficulties were experienced. This is not to suggest that procedures are a replacement for supervision. They are really only an aide and a guide to doing a specific task. They do not in themselves cause the job to be done, nor done on time, nor to be coordinated with other jobs.

Nor is the preparation of procedures a one-shot task. If it is to be effective, it must be continuing, and the lessons learned must be incorporated in each succeeding procedural update. For instance, consider the exhibits which accompany each procedure. At first it was thought that the text could use these as references, but not so. People do not like to thumb thru pages to look at the referenced exhibit. Also, a little self-examination by the analysts showed that it was difficult to actually locate data fields referred to by the procedure on the exhibits. So, we went to spot art illustrating the job step. At the same time we did not eliminate the exhibits because we found that an incoming form could be more readily identified against the procedure if there was a picture of it, rather than just a reference by name and number.

There was a very important fall-out from the preparation of procedures which had not been considered when the task began. A great amount of attention is normally made to the documentation of the overall system, to the programs in the system, and to the computer operations necessary to run the system. When we completed the task, we had not only documentation of this type, but equally detailed information on the administrative tasks which were necessary for complete system operation. The existence of this type of documentation is essential to the continuation of an on-going system upon the transfer of responsibility for operation.

The PLAYSCRIPT procedures were found to be a very effective method of directing and controlling manual editing of input data. Evidence of this is that 85% of the total number of errors in data reported from all sources were discovered during the manual editing process. Contributing to the overall effectiveness of the system reporting was that the editors were able to correct 85% of all data errors they encountered using the techniques embodied in the procedures, with the remaining 15% having to be submitted to higher levels of technical knowledge for corrective action. The final conclusion of system managers was that the effort expended in producing the procedures paid off handsomely, significantly increasing the accuracy of reports, and augmenting user confidence in the integrity of the system.

As one swallow does not make a summer, so the success encountered in this example does not prove that the application of formally written manual procedures will solve all similar problems. However, it should suggest to system and data bank managers another tool for utilization in the problem solving process. In system operation a great deal of thought and care is given to the provision of system and program documentation so that visibility to program complexities is available when unforeseen problems arise. Sometimes it can be equally worthwhile to provide the same thought and care to the documentation of the manual operations which precede and follow the computer processing portions of a total system complex.

Vitro SILVER SPRING LABORATORY OPERATIONS PROCEDURE

Subject: DR&D CONTROL PACKS IN VIVO, EDITING OF

Number: 02.13-2 **Project:** 69-58
Effective: 4-20-72
Supersedes: 02.13-1 **Date:** 7-28-69
Page 1 **of** 7

I. PURPOSE:

To establish the procedures for editing in vivo Control Packs and for the extraction of data for computer processing.

II. GENERAL:

In vivo Control Packs, consisting of a Check Sheet for Editing Control Packs, a Screening Control Record, and the required number of Screening Test Records to back up the number of tests being reported on the Control Record, are used to report in vivo test results. For each experiment, the identifying information and test results for the control group are recorded on the Control Record and for the test group on the Test Record. Data is keypunched from the source documents and used as input media to the Master Data Bank to generate a Test Record on magnetic tape for each test, and as a means of disseminating test results to suppliers, screeners, and staff. These records of testing are microfilmed and indexed for future retrieval.

III. EXHIBITS:

A. Check Sheet for Editing Control Packs, Form 1324 (Rev. 9/71)
B. Screening Control Record, Form 1157 (Rev. 1/72)
C. Screening Test Record - In Vivo, Form 1158-1 (Rev. 3/72)
D. Screening Test Record - In Vivo - Survival Systems Only, Form 1158-2 (Rev. 2/72)
E. Data Problem Resolution Form, Form 1374 (10/70)
F. EAM Section Work Sheet, DP-6 (8/68)
G. Daily Control Log, DP-76 (9-69)
H. Week Number Matrix
I. Numeric Data to Alphabetic and Alphanumeric Code Changes

J. M. O'Connor
Systems Procedures & Operations

D Roha
Chemotherapy Project Manager

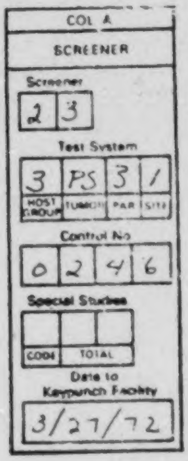
Covill

IV. PROCEDURE:

Responsibility	Action
Data Processing Assistant	1. Receives Control Packs from Project Point-of-Entry.

2. Checks Control Packs to ascertain if there are any being returned with a Data Problem Resolution Form attached. Separates these and:
    a. Detaches one copy of Data Problem Resolution Form and matches with suspense file. Destroys suspense copy and files completed copy.
    b. If punched cards were received from screener, pulls these cards from suspense and corrects.
    c. Processes these Control Packs to Step 10.

3 Edits Check Sheet against Control Record and Test Records as follows:
    a. Checks Column A of Check Sheet (1324) to see that SCREENER, TEST SYSTEM, CONTROL NUMBER, and SPECIAL STUDIES (if applicable) data fields are completed. Verifies that data in Column A of Check Sheet



COL A

SCREENER

Screener
2 3

Test System

HOST GROUP	TUMOR	PAR	SITE
3	PS	3	1

Control No
0 2 4 6

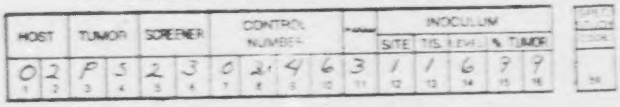
Special Studies

CODE	TOTAL	

Date to Keypunch Facility
3/27/72

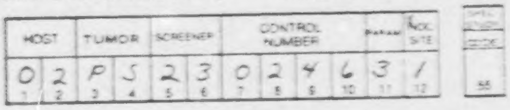
Data Processing
Assistant (cont'd.)

is in agreement with corresponding data on Control Record (1157)



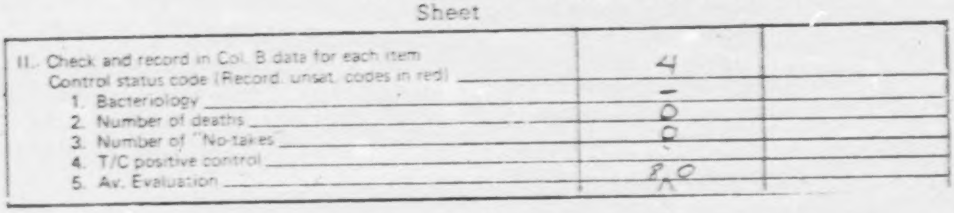
HOST		TUMOR		SCREENER		CONTROL NUMBER					INOCULUM				
											SITE	TIS. LEVEL		% TUMOR	
0	2	P	S	2	3	0	2	4	6	3	1	1	6	9	9

and Test Records (1158-1 and 1158-2)



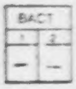
HOST		TUMOR		SCREENER		CONTROL NUMBER					INOC. SITE
0	2	P	S	2	3	0	2	4	6	3	1

b. Verifies that data in Section II of Check Sheet

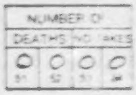


II. Check and record in Col. B data for each item Control status code (Record. unsat. codes in red)	4	
1. Bacteriology	–	
2. Number of deaths	0	
3. Number of "No-takes"	0	
4. T/C positive control		
5. Av. Evaluation	8^0	

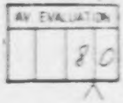
corresponds with corresponding data fields of Control Record (1157)



BACT	
1	2
–	–

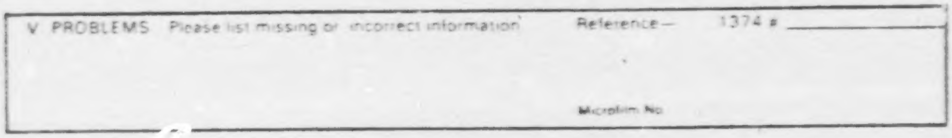


NUMBER OF			
DEATHS		NO TAKES	
0	0	0	0



AV. EVALUATION			
		8	0

Records any discrepancies in Item V. PROBLEMS, of Check Sheet.



V. PROBLEMS Please list missing or incorrect information Reference— 1374 # ______

Microfilm No.

c. If more than one page to Control Record, writes "DO NOT KEYPUNCH" on additional pages. (All pages get microfilmed.)

Data Processing Assistant (cont'd.)

4. Verifies that all data fields on the Control and Test Records requiring data are complete (makes all corrections in RED ballpoint pen). Exceptions:
   a. Negative % T/C and Positive Control NSC Number on the Control Record may be blank.
   b. If a survival system [parameter = 2, 3], Tumor Evaluation on the Control and Test Record may be blank.
   c. If "AA" Tumor, Parameter, Inoculum, Vehicle, Route, Tumor Subline, Number of Animals, Deaths, No-Takes, Cures, Control Status Code, Standard Deviation, Body Weight Change, and Average Evaluation may be blank on all records.
   d. If Control Status Code equals five or the Test Status Code equals 33 or 34, Vehicle, Route, Cures, No-Takes, Body Weight Change, Standard Deviation, and Average Evaluation may be blank on all records.
   e. Sample No. may be blank on Test Records.

5. Checks Survivors and Comments fields on Control and Test Records for screener's annotations, indicating that an animal or animals were missing.
   a. If no animals are missing - proceeds to Step 8.
   b. If animals are missing - paper clips the appropriate page or pages and forwards the Control Pack to the Quality Assurance Officer for review.

Quality Assurance Officer

6. Verifies and makes corrections if necessary to the following data fields:
   a. The Number of Animals must equal the Number of Initial Day Animals if prior to the first treatment day an animal is missing or a death occurs.

b. Toxicity Day Weights: If a mouse dies or is missing on this day, verifies that the total weights are for the remaining mice and not the original number of mice. (If a mouse dies or is missing after it is weighed, the average weight must be subtracted from the total weight batch weighed.)

7. Returns Control Pack to Data Processing Assistant.

Data Processing Assistant

8. If problems are encountered in editing:
   a. Notes them in Item V. PROBLEMS of Check Sheet.
   b. If necessary, prepares Data Problem Resolution Form in five (5) copies, pulls one and places in suspense, attaches other forms to Control Pack.
   c. Carries Control Packs with Data Problem Resolution Forms attached to Project Point-of-Entry.
   d. Logs out Data Problem Resolution Form to Program Analysis Branch on Daily Control Log.

Data Processing Assistant

9. For those Control Packs which have been received with punched cards and have passed editing, proceeds to Step 27.

10. If Control Packs with cards punched by screener have errors in cards, notes in Item V. PROBLEMS of Check Sheet. If errors cannot be resolved internally, proceeds to Step 6.

11. Prepares EAM Section Work Sheet for keypunching for Control Packs received without punched cards which have passed editing.

12. Logs out Control Packs to keypunching.

13. Carries EAM Section Work Sheet and Control Packs to Keypunch Section.

Keypunch Control Clerk

14. Receives Control Packs and EAM Section Work Sheet.

15. Logs EAM Section Work Sheet into the Data Processing Log Sheet:
    a. Originator
    b. Log Number
    c. Job/Sub-job Number
    d. Department
    e. Time In
    f. Approximate Number of Cards to be Punched.

16. Enters next consecutive Log Number from sheet onto DP-6.

17. Gives package of work to Keypunch Supervisor.

Keypunch Supervisor

18. Assigns work to Keypunch/Verifier Operators.

Keypunch/Verifier Operators

19. Punches data into cards and verifies in accordance with Keypunch Instructions 02.01-1 and 02.02-1.

20. Forwards cards and input data sheets to Keypunch Control Clerk.

Keypunch Control Clerk

21. Logs completed work in the Data Processing Log Sheet.
    a. If done same day - enters time out in original entry.
    b. If not done same day - enters new line of data using previously assigned Log Number.

Keypunch Control Clerk (cont'd.)

22. Forwards completed work to Keypunch Supervisor.

Keypunch Supervisor

23. Verifies that punching and verifying have been completed in accordance with Keypunch Instructions 02.01-1 and 02.02-1.

24. Places work on job pick-up table.

Data Processing Assistant

25. Picks up completed work at Keypunch Section.

26. Pulls EAM Section Work Sheet and places in completed box on job pick-up table.

NIH Project Point-of-Entry

27. Logs in completed keypunch work on Daily Control Log. Turns work over to Data Processing Assistant.

Data Processing Assistant

28. Edits cards as follows:
    a. Checks that there is one card with a "1" in Card Column 80 and one card with a "2" in Card Column 80 for each Control Pack.
    b. Checks cards with "3" in Card Column 80 to verify that there is one for each Test Record.

29. Places control and detail cards in "CURRENT WEEK PROCESSING DRAWER" for daily Pre-Validity Run.

30. Places Control Packs in hold for weekly microfilming.

J. M. O'Connor, Analyst

EXHIBIT A.

Instructions for use:

1. See Inst. 117A (9/71).
2. Screening Supervisor—on Day Final, prepare 1324 for each In Vivo control pack. Edit items I and II and complete Columns A and B. Attach Special Study requests, and original and hardback copies of 1324 to the final Day Screening Control and Test Records and forward to the keypunch facility or Data Processing (DPC) as directed. The Screener copy may be retained or discarded at the discretion of the screening laboratory.
3. Keypunch Facility Complete item III.
4. Data Processing Contractor—Edit items 1A and II. Record Problems under Item V. Complete column C. Process Forward copies
5. Chemotherapy—Review PROBLEMS. Distribute copies.

CHECK SHEET FOR EDITING CONTROL PACKS

COL A SCREENER	Please check the items listed below for accuracy Initial and date in the appropriate column to the right	COL B SCREENER	COL C DATA PROCESSING CONTRACTOR
Screener 2 3 Test System 3 PS 3 1 HOST GROUP / TUMOR / PAR / SITE Control No 0 2 4 6 Special Studies CODE / TOTAL Date to Keypunch Facility 3/27/72	1. A. 1. CONTROL RECORD-All information filled in and correct 2. NSC numbers (MUST AGREE WITH TEST RECORDS) 3. Calculation check a. Av. Evaluation b. Body Weight Change: Final-Initial Day 4. Positive Control a. Total Tumor Evaluation b. Average Evaluation c. Body Weight Change d. %TC 5. Cell Culture Only a. C_0 b. C c. C/C_0 d. COD_0 e. COD	m.m. 3/23/72	Date Rec'd 3/30/72 Punch Cards ☒ Yes ☐ No Batch No 912-66 Edited by N S Date 3/30/72
	B. All Tests-all information filled in and correct. 1. Vehicle 2. Route 3. Treatment Schedule (Interval through Dose) 4. Number of Survivors-end/start 5. Calculation check a. Totals-Initial and Final Day animal weights b. Total Tumor Evaluation c. Average Evaluation d. Body Weight Change e. %TC 6. Test status code	B.9. 3/24/72	

II. Check and record in Col. B data for each item Control status code (Record. unsat. codes in red)	4.	
1. Bacteriology	—	
2. Number of deaths	0	
3. Number of "No-takes"	0	
4. T/C positive control		
5. Av. Evaluation	8.0	

III. Date rec'd 3/27/72 Keypunched by K.B. Verified by B.B. Date to DPC 3/29/72

IV. Chemotherapy

A PRA

1 Date Rec'd ______ Problems Rev'd by ______ Date ______

2 Date to project officer ______

B DEB

1 Date to screener ______ by ______

V. PROBLEMS Please list missing or incorrect information Reference — 1374 # ______

Microfilm No

1324 (Rev. 9/71)

DATA PROCESSING CONTRACTOR / D R & D

EXHIBIT B.

SCREENING CONTROL RECORD

HOST	TUMOR	SCREENER	CONTROL NUMBER		INOCULUM SITE	INOCULUM TIS.	INOCULUM LEVEL	INOCULUM % TUMOR	DATE ON YEAR	DATE ON MONTH	DATE ON DAY	VEH.	RT	NO OF MATERIALS	NUMBER OF TESTS TOTAL	NUMBER OF TESTS SYNTH	NUMBER OF TESTS PLANTS
0 2	P S	2 3	0 2 4 6	3	1	1	6	9 9	7 2	0 2	2 3	2	1	0 7	2 5	2 5	0 0
1 2	3 4	5 6	7 8 9 10	11	12	13	14	15 16	17 18	19 20	21 22	23	24	25 26	27 28	29 30	31 32

NO. OF TESTS FERM. PROD.	NO. OF TESTS ANIM PRODS.	TUMOR SUBLINE LINE	TUMOR SUBLINE CODE	TUMOR SUBLINE GEN	ANIMALS SEX	ANIMALS SOURCE	ANIMALS NUMBER		FINAL EVAL DAY	NUMBER OF DEATHS	NUMBER OF NO-TAKES	NEGATIVE %T/C	C S C	SPEC STOCK CODE		POSITIVE CONTROL NSC NUMBER
0 0	0 0	0 4	Y	1 4	F	3 5	30		3 0	0 0	0 0		4			
33 34	35 36	37 38	39	40 41	42	43 44	45 46		49 50	51 52	53 54	55 56 57	58	59		74 75 76 77 78 79

SURVIVORS

DAY	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
NO		30						26	9	2	1	0													
DAY	25	26	27	28	29	30																			
NO																									

	NSC NUMBERS	NO	ANIMAL WEIGHTS INITIAL DAY (DAY 1)	ANIMAL WEIGHTS TOXICITY EVAL (DAY 05) 47 48	TUMOR EVALUATION	NO	ANIMAL WEIGHTS INITIAL DAY 1	ANIMAL WEIGHTS TOXICITY EVAL DAY 5	TUMOR EVALUATION
1	27639	1	1140	1250		21			
2	"	2	1160	1260		22			
3	"	3	1160	1265		23			
4	"	4	1120	1290		24			
5	29408	5	1180	1260		25			
6	"	6				26			
7	"	7				27			
8	"	8				28			
9	"	9				29			
10	30625	10				30			
11	"	11				31			
12	"	12				32			
13	"	13				33			
14	38406	14				34			
15	"	15				35			
16	"	16				36			
17	39112	17				37			
18	"	18				38			
19	"	19				39			
20	39274	20				40			
21	"								
22	"								
23	45024								
24	"								
25									

COMMENTS:

TOTAL	0	5	7	6	0	0	6	3	2	5				
	60	61	62	63	64	65	66	67	68	69	70	71	72	73
AV			1	9	2			2	1	1				

BACT 1	BACT 2	STAND. DEVIATION		PROBIT ERROR	AV. EVALUATION
—	—	10	±	+ 019	80

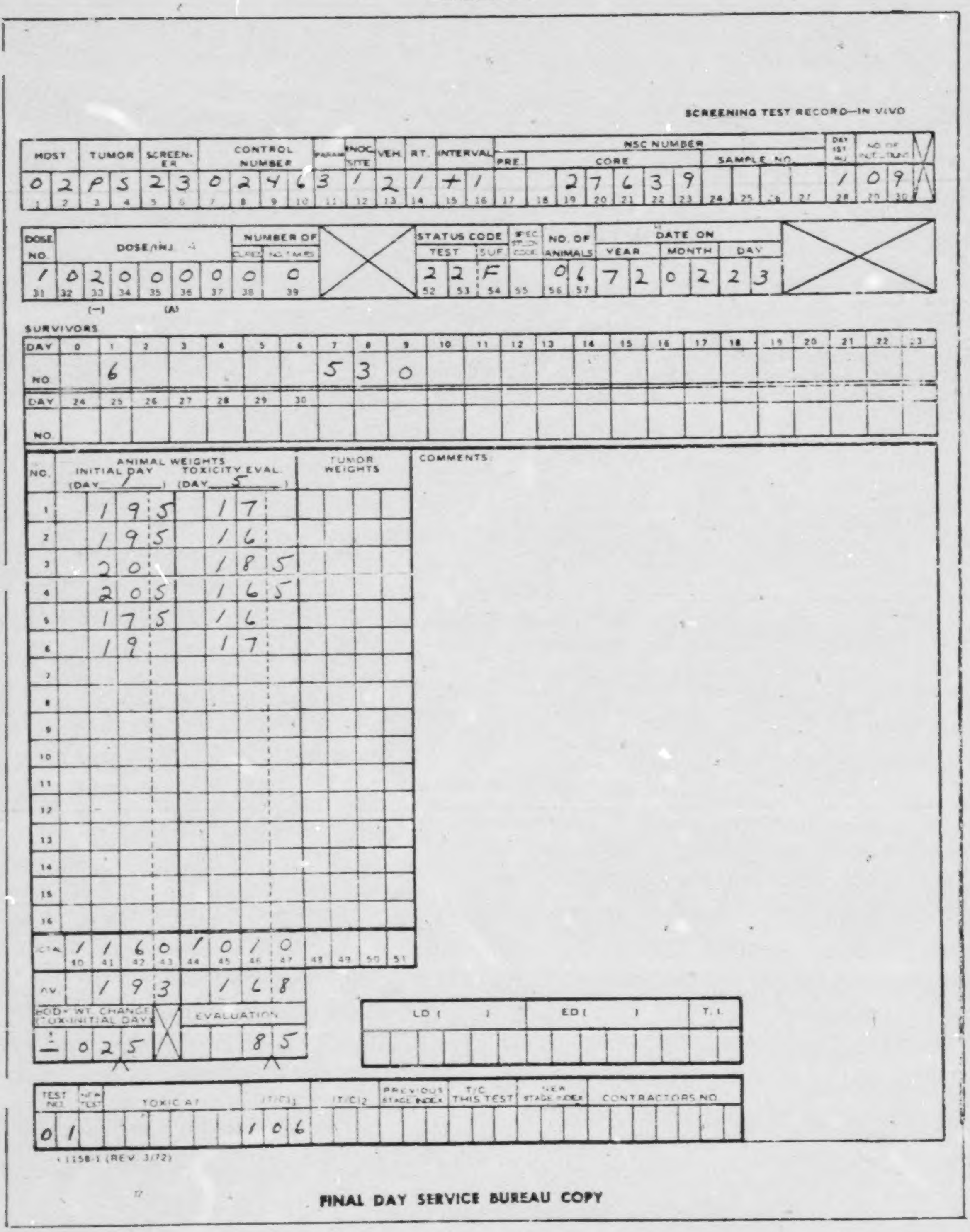
CONTROL

1157 (REV. 1/72)

DATA PROCESSING CONTRACTOR/DR&D

Covill

EXHIBIT C.



SCREENING TEST RECORD—IN VIVO

HOST		TUMOR		SCREENER		CONTROL NUMBER				PARAM	INOC SITE	VEH.	RT.	INTERVAL		NSC NUMBER PRE.		CORE					SAMPLE NO.				DAY 1ST INJ	NO OF INJECTIONS	
0	2	P	S	2	3	0	2	4	6	3	1	2	1	+	1			2	7	6	3	9					1	0	9
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

DOSE NO.	DOSE/INJ						NUMBER OF CURED	NO. TAKES	STATUS CODE TEST		SUF.	SPEC. STUDY CODE	NO. OF ANIMALS		DATE ON YEAR		MONTH		DAY	
1	0	2	0	0	0	0	0	0	2	2	F		0	6	7	2	0	2	2	3
31	32	33	34	35	36	37	38	39	52	53	54	55	56	57						
		(−)		(A)																

SURVIVORS

DAY	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
NO		6						5	3	0														
DAY	24	25	26	27	28	29	30																	
NO																								

NO.	ANIMAL WEIGHTS INITIAL DAY (DAY 1)	TOXICITY EVAL. (DAY 5)	TUMOR WEIGHTS
1	195	17	
2	195	16	
3	20	185	
4	205	165	
5	175	16	
6	19	17	
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
TOTAL	1160 (40–43)	1010 (44–47)	(48–51)
AV.	193	168	

COMMENTS:

BODY WT. CHANGE (TOX-INITIAL DAY)	EVALUATION
± 025	85

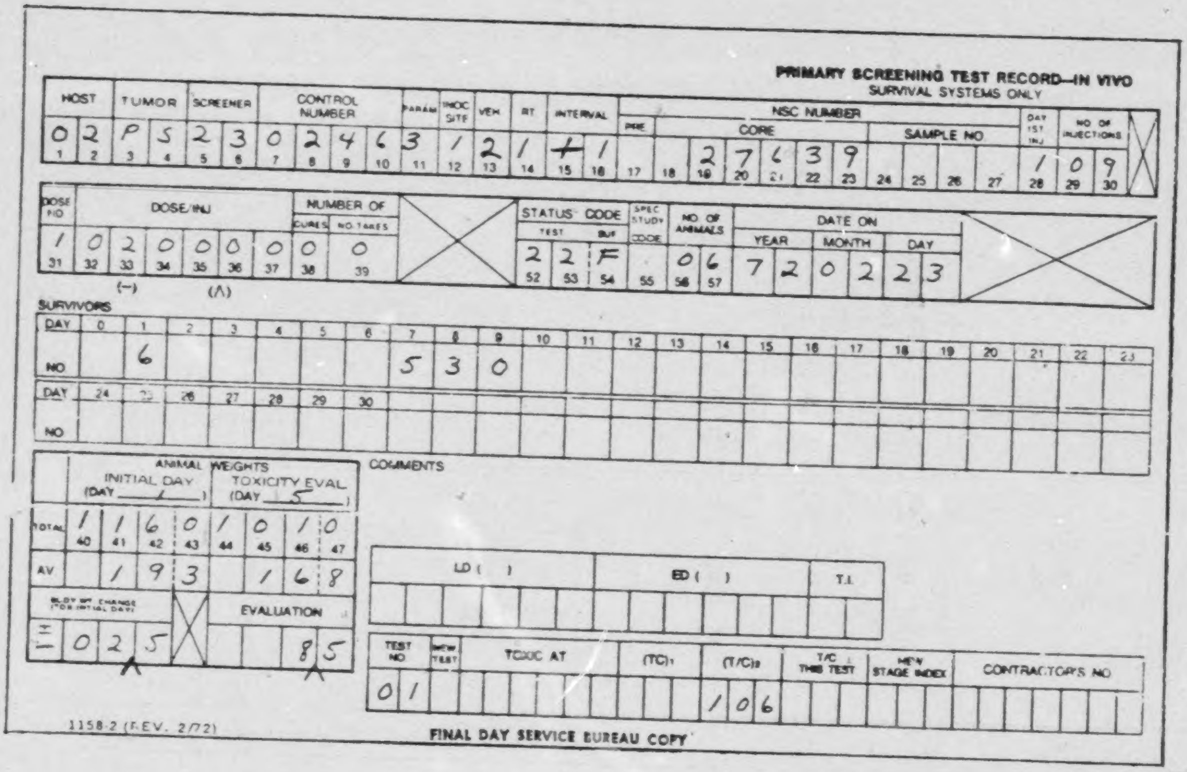
LD ( )	ED ( )	T. I.

TEST NO.	NEW TEST	TOXIC AT	$(T/C)_1$	$(T/C)_2$	PREVIOUS STAGE INDEX	T/C THIS TEST	NEW STAGE INDEX	CONTRACTORS NO.
01			106					

1158-1 (REV. 3/72)

FINAL DAY SERVICE BUREAU COPY

EXHIBIT D.



PRIMARY SCREENING TEST RECORD—IN VIVO
SURVIVAL SYSTEMS ONLY

HOST		TUMOR		SCREENER		CONTROL NUMBER				PARAM	INOC SITE	VEH	RT	INTERVAL		NSC NUMBER PRE		CORE					SAMPLE NO				DAY 1ST INJ	NO OF INJECTIONS	
0	2	P	S	2	3	0	2	4	6	3	1	2	1	+	1			2	7	6	3	9					1	0	9
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

DOSE NO	DOSE/INJ						NUMBER OF CURES	NUMBER OF NO TAKES	
1	0	2	0	0	0	0	0	0	
31	32	33	34	35	36	37	38	39	
		(−)		(Λ)					

STATUS CODE TEST		STATUS CODE BUF	SPEC STUDY CODE	NO OF ANIMALS		DATE ON YEAR		DATE ON MONTH		DATE ON DAY	
2	2	F		0	6	7	2	0	2	2	3
52	53	54	55	56	57						

SURVIVORS

DAY	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
NO		6						5	3	0														

DAY	24	25	26	27	28	29	30
NO							

ANIMAL WEIGHTS

	INITIAL DAY (DAY 1)				TOXICITY EVAL (DAY 5)			
TOTAL	1	1	6	0	1	0	1	0
	40	41	42	43	44	45	46	47
AV		1	9	3		1	6	8

BLDY WT CHANGE (FOR INITIAL DAY)					EVALUATION			
±	0	2	5				8	5
		^					^	

COMMENTS

LD ( )	ED ( )	T.I.

TEST NO		NEW TEST	TOXIC AT	(TC)₁	(T/C)₂			T/C THIS TEST	NEW STAGE INDEX	CONTRACTOR'S NO
0	1				1	0	6			

1158-2 (REV. 2/72)

FINAL DAY SERVICE BUREAU COPY

EXHIBIT E.

DATA PROBLEM RESOLUTION FORM

TO:	FROM: DPC	DATE: 9 Mar. 72	ACC# V-160
DOCUMENT In Vivo Control Pack Control #1956			MICROFILM#
PROBLEM			SCREENER 09

"HOST" blank. Need "HOST" data. See Paragraph 4, Procedure 02.13-2.

INITIALS R.H.P.

ANSWER

1374 (10/70) ORIGINATOR'S COMPLETED COPY INITIALS ________ DATE ________

39

Covill

EXHIBIT F.

DP-A (Rev. 8-68)

Vitro LABORATORIES
SILVER SPRING LABORATORY

M SECTION WORK SHEET

DESCRIPTION NIH	WORK REQUEST #		JOB CHARGE #	01661.10014
ORIGINATOR Jane Doe	EXT. 2181	SECTION CP-11	SECURITY CLASSIFICATION	☐ SECRET ☐ VITRO CONFIDENTIAL ☐ CONFIDENTIAL ☒ UNCLASSIFIED
SIGNATURE Jane Doe	DATE/TIME SUBMITTED 17 Mar. 72	DATE REQUIRED 17 Mar. 72 12 NOON	DATE/TIME COMPLETED	

FUNCTION | STEP INFORMATION | See NH 02.01-1, 02.02-2 | INSTRUCTIONS | Use 029

KEY PUNCH

STEP
INITIAL

DATE/TIME RECEIVED	DATE/TIME COMPLETED	NO. PAGES ATTACHED	BEGIN COMMENT I.D.	IDENTIFICATION PUNCH	IDENTIFICATION COLUMNS	PROGRAM NO.	PUNCHED BY
		NO. CARDS SUBMITTED 150.				VERIFIED BY	CC & ERROR COUNT NO.

SORTING

STEP
INITIAL

DATE/TIME		NO.	FIELD	COLUMNS		TYPE	
RECEIVED	COMPLETED			FROM	TO	ALPHA	NUMERIC
		1.					
		2.					
		3.					

REPRODUCING & INTERPRETING

STEP
INITIAL

☐ REPRODUCE ☐ INTERPRET ☐ 80-80 (514) ☐ 60-60 (557)

DATE/TIME		OLD DECK IDENTIFICATION	CARD COLUMNS (FROM)	514 GANG PUNCH INFO.	514 REPRO CARD COL. TO	557 PRINT POSITION	557 LINE POSITI
RECEIVED	COMPLETED						

TABULATING

STEP
INITIAL

DATE/TIME		BOARD	PAPER		SPACING	CARRIAGE CONTROL
RECEIVED	COMPLETED	☐ 80-80	☐ FORM NO. ________		☐ 1	☐ STANDARD
		☐ BOARD NO.	☐ STOCK	☐ TWO PART	☐ 2	☐ SPECIAL *(See Instr.)*
		☐ SPECIAL *(See Instr.)*	☐ NARROW	☐ FOUR PART	☐ 3	☐ TAPE NO.
			☐ ONE PART	☐ SIX PART		
		ALTERATION SWITCH(ES) ON			☐ BURST	☐ TABULATE

COLLATE

STEP
INITIAL

DATE/TIME		CARD TYPE		CONTROL COLUMNS		OPERATION
RECEIVED	COMPLETED	PRIMARY		MAJOR		☐ SEQUENCE CHECK
						☐ MATCH
		SECONDARY		INTERMEDIATE		☐ SELECT
				MINOR		☐ MERGE

SPECIAL

STEP
INITIAL

52

Covill

EXHIBIT G.

Vitro LABORATORIES
SILVER SPRING LABORATORY

DAILY CONTROL LOG

DATE RECEIVED	BATCH NO	SCREENER	TEST SYSTEM	CONTROL	MICROFILM NO	KEYPUNCH TO	KEYPUNCH FROM	PROB RESO TO	PROB RESO FROM	DATE TO MIM	COMMENTS
1-7-72	6	09	3LE21	2250	002085					1-21-72	wc
				2255	090						
				2256	091						
			↓	2257	092						
			3PS31	212	103						
				213	104						
				214	105						
				215	106						
				217	107						
			↓	218	108						
		↓	5WA16	708	111						↓
		02	3LE21	4212	006	1-9	1-12				
				4221	008						
			↓	4236	013						
			3PS31	593	016						
				597	017						
		↓	↓	604	020	↓	↓				
		08	3LE21	4142	073						wc
				4143	074						
			↓	4161	077						
		↓	3PS31	34	078						
		09	3LE21	2251	086						
				2252	087						
				2253	088						
				2254	089						
			↓	2260	095						↓
		↓	~~3PS31~~	~~216~~							
		10	3LE21	1709	117						wc A28
				1721	118						A24
				1723	119						A24
	↓	↓	↓	1734	120						↓ A24
~~1-12-72~~	~~7~~	~~01~~	~~3AA4~~	~~210~~							~~B50~~
	↓	↓	3LE21	1705							B56
	7	02	3LE21	4223	009	1-12	1-13				
			↓	4234	011						
			3PS31	606	022						
			↓	614	025						
↓	↓	↓	5WA16	1499	↓ 030	↓	↓			↓	

EXHIBIT H.

WEEK NUMBER MATRIX, FISCAL YEAR 1972

The date for each run is taken from this matrix as YY=1971 or 1972, MM=Month and DD=final date of given run period as 710708 for first given period in July.

Run Week #	Inclusive Dates	Run Week #	Inclusive Dates
127	July 1 - July 8	201	December 28 - January 5
128	July 8 - July 15	202	January 5 - January 12
129	July 15 - July 22	203	January 12 - January 19
130	July 22 - July 29	204	January 19 - January 26
131	July 29 - August 5	205	January 26 - February 2
132	August 5 - August 12	206	February 2 - February 9
133	August 12 - August 19	207	February 9 - February 16
134	August 19 - August 25	208	February 16 - February 23
135	August 25 - September 1	209	February 23 - March 1
136	September 1 - September 8	210	March 1 - March 8
137	September 8 - September 15	211	March 8 - March 15
138	September 15 - September 22	212	March 15 - March 22
139	September 22 - September 29	213	March 22 - March 29
140	September 29 - October 6	214	March 29 - April 5
141	October 6 - October 13	215	April 5 - April 12
142	October 13 - October 20	216	April 12 - April 19
143	October 20 - October 27	217	April 19 - April 26
144	October 27 - November 3	218	April 26 - May 3
145	November 3 - November 10	219	May 3 - May 10
146	November 10 - November 17	220	May 10 - May 17
147	November 17 - November 24 *	221	May 17 - May 24
148	November 24 - December 1 *	222	May 24 - May 31
149	December 1 - December 8	223	May 31 - June 7
150	December 8 - December 15	224	June 7 - June 14
151	December 15 - December 22 *	225	June 14 - June 21
152	December 22 - December 28 *	226	June 21 - June 28

NOTE * Normally, the volume of data is small around holidays so input for Weeks 147 and 148 is combined; the same applies to Weeks 151 and 152.

EXHIBIT I.

NUMERIC DATA TO ALPHABETIC AND ALPHANUMERIC CODE CHANGES

When a two character numeric code is entered into the DAY 1ST INJ., NO-TAKES or CURES fields, it will be corrected by entering an alphabetic code as:

10 = A
11 = B
12 = C
13 = D
14 = E, etc.
thru
35 = Z

When a two character numeric code is entered into the DOSE NO. field, it will be corrected by entering an alphanumeric code as:

10 = +
11 = A
12 = B
13 = C
14 = D, etc.
thru
36 = Z

An Adaptive File Management Systems

Dennis Lee Dance*
University of Arkansas
Little Rock, Arkansas

Udo W. Pooch*
Texas A&M University
College Station, Texas

A program module is described defining an interface between an online - information system and the Input/Output Control System of the computer system. Programs belonging to this module are grouped by the function they perform: buffering, item relocation, compression, and dynamic priority assignment. The interface is adaptive in nature by physically reorganizing the File Structure based on usage statistics. Records are physically assigned to priority areas to reduce system I/O. The results of the reorganization is to construct working set files, a subset of the original file structure, having a substantial portion of all file activity. This working set is maintained in core via buffering, thereby reducing I/O overhead.

Key words: Adaptive file management; hierarchial file structure; file working set; self-organizing data sets.

1. Introduction

With the advent of time-sharing computer systems many of the functions and services provided by libraries can be automated. Storage devices such as disks contain information, maintenance programs keep the information current, and query languages provide a means of retrieval. If a system can perform these tasks, it is known as an Information Storage and Retrieval System (ISRS) [17,11,12]. If in addition it can detect trends in the data, it is called a Management Information System (MIS) [9,16].

An Online Data Compression System (ODCS) must be an invisible interface between the information retrieval system and the Input/Output Control System (IOCS) of the computer operating system. This paper describes a set of programs which are analogous to a demand-paged memory management operating system. These programs are classified into three main categories: compression/decompression, buffering, and relocation. The compression/decompression routines reduce the physical storage requirements to contain a particular set of data items. The buffering routines maintain several physical records in main memory to reduce the I/O traffic. The relocation routines provide the computer with the ability to "reorganize" those items which are requested most often, and to physically reorganize those times for easy, fast retrieval, much analogous to the working set notion of virtual memory [6].

---

* Computer Scientists

2. Design Philosophy

An ODCS must be designed to relieve the ISRS/MIS of the burden of physical file structure (list, tree, etc.) Allowing the ISRS/MIS to maintain logical items in logical files simplifies the ISRS/MIS program complexity. Another interface function is to reduce the volume of space required to contain the data file, and thereby freeing storage devices for either other use or lower costs for fewer devices. The compression, applied on a record-by record basis rather than on the entire file, results in fewer I/O requests for a given amount of data because more data is transferred per request. The interface must also attempt to reduce I/O requests by physically grouping items based on their usage frequency. Thus one I/O access may retrieve several items having a high probability of being requested next.

Whenever a large number of data items must be maintained, some fast look-up or address generator mechanism must be used to locate the desired information. Hashing algorithms, illustrated in figure 1, map logical keys into uniformly distributed equivalence classes [5,14,18]. Unfortunately, these relations do not always provide unique physical addresses. Intrarecord collisions occur whenever the mapping generates the same address more than once, and the item can be stored at the generated location; a pointer locates the item within the record. Inter-record collisions occur whenever no more space is available at the target address and the item must be chained to another record.

Even though the creation of activity affinity groups can reduce the number of accesses, further reduction may be accomplished by performing buffering. This buffering is not to be confused with that performed by the Input/Output Control System of the computer. Instead, this buffering should be designed to retain in memory for a "reasonable" length of time those physical records most frequently and last requested (least recently used algorithm). This should allow additional logical items to be requested but with fewer physical accessions. The relationship and organization of the new ISRS/MIS is shown in figure 2.

3. System Implementation

The implementation of the interface requires several tables to define interface parameters, to locate data sets in the data base, to define and describe data sets in the data base, to reduce excessive I/O activity, and to locate individual items in a physical record

A single table, System Map (SYSMAP), initially defines the interface parameters for system generation. These parameters include the current time or date, the initial and the incremental size of the Collision Relocation List (CRL), the sizes of n-tuples used to build other tables, threshold values for changing item priorities, and threshold values for data set reorganization. Several of the parameters, discussed later, in SYSMAP can be modified during program execution to allow the interface to adapt to its environment.

The definition of the tables used to locate the data sets depends on both system parameters and user specified parameters. The tables used for data set location include the Available Space for Information Maps List (AVSIML), the Data Set Information Map List (DSIML), and the Data Set Information Map (DSIM). Tables associated with each data set include the Data Set Statistics Table (DSST), the Frequent Collision List of Items (FCLSTI), the Most Frequent Item List (MFIL), the Record Information Field (RIF), the CRL, the Item Information Field (IIF), and the Item Value Area (IVA). Each of these tables is described and summarized in Table 1.

Table 1. Table Summary

Table Name	Contents	Function
SYSMAP	system parameters; table sizes; threshold values	supplies parameters to DSIM: defines n-tuples
AVSIML	pointers	locates free space for DSIML n-tuples
DSIML	pointers	locates DSIM
DSIM	data set parameters; threshold values	defines record area boundaries; provides threshold values
DSST	statistics	provides synopsis of data set activity
MFIL	pointers	locates high activity items
FCLSTI	pointers	locates high activity relocated items
RIF	counters and pointers and reserve switch	records number and types of collisions; indicates if record is reserved; locates CRL, IIF, and IVA
CRL	pointers	locates relocated items
IIF	pointers and counters	describes and locates an item
IVA	data value	compressed storage area

Some of the data set parameters are specified by the user while others are obtained from SYSMAP. User-specified parameters include the acceptable time delay for updating, retrieving, and storing items in the data set; the beginning and ending physical record addresses for the initial storage area segregating the items by type; the beginning and ending physical record addresses segregating the collision overflow area by item type; the beginning and ending physical record addresses for the reserved record area.

Figure 3 illustrates the logical table relationship for data base, as opposed to data set, maintenance. The dashed line indicates communication of parameters while the solid line indicates linking information. The two disk volumes shown in figure 3 are logical disks since they may physically be on one volume or require several volumes. Regardless of the physical size, the logical relationship remains the same. The SYSMAP supplies parameters to each DSIM which in turn contains parameters defining the boundary areas for the data records comprising a data set in the data base.

To facilitate program development, maintenance, and modification, the system was written in modular subroutines. These program modules, with the exclusion of the main subroutine ODCS, can be divided into three main classes: buffering, compression/decompression, and relocation.

The purpose of the buffering routines, GETREC and PUTREC, is to maintain several physical records in memory to reduce the physical volume of I/O traffic. Records are read from disk when demanded by the system.. Once in core, an activity counter is associated with each buffer pool allowing frequently requested records to remain in core while other less-demanded records in the buffers can be replaced. Reading from disk into memory and passing the requested physical record to the other system routines is performed by GETREC. The routine PUTREC receives data from the processing routines and stores the data either in a buffer or on disk.

The relocation routine, ANOTHR, uses both the buffering and compression/decompression routines to relocate an item from its home record to another record. Depending on the parameters in the DSIM and the subroutine call arguments, the item is relocated in either the collision overflow area or the reserved area. Additionally, whenever an item is relocated, the two tables MFIL and FCLSTI must be examined to determine if these tables contain this item so that the entries can be updated. The relocation procedure is simply a sequential scan of the records in a predefined area. The scan stops whenever a record is located that can contain the item.

3.1 System Operation

The system is first generated with a BLOCK DATA subroutine and data sets defined with both system- and user-supplied parameters. Items which can be forced by type into partitioned areas may be entered, retrieved, or deleted. Items are relocated whenever one of the following occurs: insufficient space is available at the home record; the compression time is exceeded: the item activity forced a priority change; or the item is placed into the reserved area. The IIF handles intra-record collisions while the CRL controls inter-record collisions and item relocation into the reserved area. The system automatically changes item priorities and creates reserved records, depending on the item activity. Excessive I/O activity is reduced by buffering and by using both the FCLSTI and reserved records.

3.2 Changing Item Priority

The purpose of assigning priorities to items is to dictate the amount of time required to retrieve from the data set the desired information. That is, if a priori knowledge is available concerning the demand for the items comprising the data set, the items can be assigned priorities. The higher the priority, the shorter the time acceptable to retrieve that item. For example, consider a company selling stocks and bonds. The political events of the day might easily cause a set of stocks or bonds to be traded extensively. Consequently, a dynamic priority assignment would be needed to allow the retrieval time for stock quotations to be reduced as the volume of trade increases.

Dynamic priority modification is accomplished by maintaining an update counter and retrieval counter for each item in the data set. Also a date is used to trigger the process of determining whether the item's priority is to be increased, decreased, or to remain the same. Based on SYSMAP parameters and threshold values, if an item activity is sufficiently great, the item information will be entered into the MFIL and FCLSTI reducing future system overhead. This automatic, dynamic priority classification process allows items to be grouped physically by priority, and thereby reducing future I/O overhead.

3.3 Reserved Records

The approach used in creating different organizational hierarchies is analogous to cache memory, with the last area prior to main memory being the reserved area. The most active items in the data set, regardless of priority, reside in the reserved area. If the item activity drops below a threshold in the DSIM, the item is returned to its home record as determined by its priority.

The creation of reserved records in performed by initially entering into the MFIL only the most active items. Once the MFIL is "sufficiently full", these items are then physically placed in the reserved area. To prevent excessive overhead and to allow the system and data set to reach a steady state, the MFIL is not considered for replacement until after a threshold number of requests have been made to the system. Also, the MFIL cannot be replaced unless the number of items in the MFIL is above a certain threshold. The first threshold allows those items in the reserved area to remain there until a certain number of operations have been performed, possibly to create a new set of items in greater demand. Consequently, when this threshold has been reached, if the number of items in the MFIL is above the second threshold, it is replaced. These two threshold values, as well as the threshold values for priority assignment, determine to a large extent the amount of system overhead.

3.4 Item Operation

After utilizing a modular (item dependent) hashing function to determine the physical address of the item, a determination of the required operation is performed. If the operation is to enter an item, GETREC retrieves the desired record and DECOMP decompresses it. After searching the IIF table in the record to determine that no duplicate exists, TRYCMP attempts to put the item in the retrieved record. If successful, PUTREC restores the record. If the compression attempt is unsuccessful, either a CRL for that record is constructed, if one is not present, or the existing CRL is expanded, if full. If possible, the ANOTHR subroutine relocates the item resulting in updating the CRL and in restoring the record via PUTREC. If the item cannot be relocated or the CRL cannot be expanded, appropriate return codes are set and a return to the calling program is executed. Following a successful item entry, the item is also entered into FCLSTI, if both relocated and more active than the least active entry in FCLSTI. Similarly, if the item is active enough, it is entered in the MFIL, replacing the least active entry.

After examination of the appropriate thresholds, the MFIL is replaced to create reserved records. Using the address pointers in the MFIL n-tuples, GETREC retrieves and DECOMP decompresses the record containing the item to be placed in the reserved area. If the item is in its home record, it is deleted, is relocated by ANOTHR, the home record's CRL is updated, and the home record is restored via PUTREC. If the item, when retrieved by GETREC using the address supplied by the MFIL n-tuple, is determined to be a relocated item, the item is deleted from the relocation record, is relocated in the reserved area, the home record is retrieved, the CRL entry is updated, and the home record is restored. If at anytime the creation of a reserved record is not successful (e.g. no available space), the procedure is halted and the item either is returned to the record from which it was obtained or is relocated in the nonreserved area.

If the operation is to retrieve an item, the FCLSTI is searched to obtain the physical address. If not located in the FCLSTI, the modular hashing function generates the address. In either case, GETREC retrieves the desired record, DECOMP decompresses it, and the IIF is searched for an item match. If found, the item is retrieved; if not found in the IIF, the CRL is searched for an item match to produce the relocation address. If the CRL points to the item, GETREC retrieves that record and the procedure is repeated. If the item is not locatable, return codes are set and a return to the calling program is executed. If retrieval is successful, priority change determination takes place depending on item activity, cutoff dates, and activity thresholds. If the priority is changed, the system deletes and reenters the item, updating the priority status, and frequency counters are returned to the user. After completing the retrieval, the FCLSTI and MFIL checks are performed again.

If the operation is to delete an item, without searching the FCLSTI, the modular hashing function generates the address used by GETREC to retrieve the home record decompressed by DECOMP containing the item to be deleted. If the item is located via the IIF, it is deleted; the relocation record is retrieved and the item is deleted from that record, and any records modified are restored by PUTREC. Whenever an item is deleted, all the activity and priority information is returned with the item, to facilitate re-entry if the item is to be modified. To complete deleting the item from the data set, any reference to the item in FCLSTI or MFIL is removed.

4. Performance Analysis

Twenty-five data sets were generated for test purposes. Four hundred uniformly generated transformations, either retrievals or updates, were applied to the data sets. Of the twenty five data sets, only the twenty-fifth data set was unique, in that the results are indicative of an ISRS/MIS without an interface.

The summary table of I/O activity, appearing in Table 2, describes the accession perturbations, resulting from parametric variations, for record and item buffering. When records are buffered, the most active records are maintained in memory; whereas, when items are buffered, the most frequently requested items are kept in memory.

The ODCS was designed to improve overall performance. One of the expected improvements was a reduction in I/O traffic by maintaining a list, FCLSTI, of the most frequently referenced relocatable items. The amount of reduction ranged from twelve to ninety-eight I/O accessions. The lower values were the result of changing the priority of an item on a collision chain and storing it on another record with no subsequent relocation. The higher values resulted from having a large number of items entering the reserved area and/or having the items on a collision chain. In either case, the use of the FCLSTI did reduce the number of I/O accessions.

Table 2. I/O Accession Summary

								Buffer Records Reserved Records Collisions				Buffer Items No Reserved Records No Collisions	
								Result 1		Result 2		Result 3	
								Updates+Retrievals		Updates+Retrievals		400 Retrievals	
Data Set	NITMFL	MFLNTH	UPTHRS	RETHRS	UDTHRS	RDTHRS	SDTHRS	Accesses	% of 635	Accesses	% of 600	Accesses	% of 400
1	0.5	10	1	1	1	1	1	2072	326	480	80	92	23
2	0.5	10	1	1	1	1	100	1829	288	474	79	74	19
3	0.5	10	1	1	1	100	1	1928	302	472	79	72	18
4	0.5	10	1	1	100	100	1	1776	280	474	79	74	19
5	0.5	10	1	20	1	1	1	780	123	523	87	121	30
6	0.5	10	20	20	1	1	1	484	76	600	100	400	100
7	0.5	50	1	1	1	1	1	1196	188	633	105	203	51
8	0.5	50	1	1	1	1	100	1073	169	628	105	228	57
9	0.5	50	1	1	1	100	1	1158	182	643	107	243	61
10	0.5	50	1	1	100	100	1	1098	173	616	103	200	50
11	0.5	50	1	20	1	1	1	727	114	513	86	203	51
12	0.5	50	20	20	1	1	1	576	91	600	100	400	100
13	0.9	50	1	1	20	20	20	1006	158	633	105	233	58
14	0.9	50	1	1	20	20	40	962	151	608	101	208	52
15	0.9	50	1	1	30	30	30	960	151	683	114	205	51
16	0.9	50	1	1	50	50	50	1158	182	637	106	201	50
17	0.9	50	1	20	20	20	20	645	101	508	85	158	[illegible]
18	0.9	50	20	20	20	20	20	564	88	600	100	400	100
19	0.9	50	1	1	1	1	1	1028	162	643	107	243	61
20	0.9	50	1	1	1	1	100	1092	172	623	104	223	56
21	0.9	50	1	1	1	100	1	974	153	643	107	[illegible]	[illegible]
22	0.9	50	1	1	100	100	1	922	145	648	108	249	62
23	0.9	50	1	20	1	1	1	646	102	509	85	108	27
24	0.9	50	20	20	1	1	1	440	69	600	100	402	100
25								635	100	600	100	400	100

Another improvement was expected from the automatic priority classification of items based on usage statistics. When records are buffered and no reserved records are created (e.g. data sets five, eleven, and seventeen), little or no advantage was apparent, because records were not kept dense with items as a result of the sequential scan used to locate free record space. If items were buffered, approximately a twenty percent reduction in I/O activity (relative to data set twenty-five) could be produced. The reason for this reduction is because items that have a high request frequency would be stored in memory to create a working set data set. A trivial solution to determine the items to buffer in memory would be to set a low activity threshold so that high activity items would displace low activity items, resulting in a flurry of double accessions prior to reaching a steady state. An alternative to constructing these core resident records would be to require a sufficient number of disk accessions to each item to indicate its classification. Regardless of the priority determination procedure, using low threshold values presupposes a priori knowledge about the data set activity.

A major benefit of the interface is a reduction in the storage space required to store the data. Of the many compression techniques available, a simple algorithm was selected solely for demonstration purposes [2-4,8,10 13,15,17]. This word redundancy reduction procedure produced approximately eighteen percent savings in space. More advanced techniques [13] should produce even greater compression.

The parameters modified during the testing procedure were the activity percentage for item entry into the MFIL (NITMFL); the number of operations occuring just prior to attempting reserved record creation (MFLNTH); the update (UPTHRS) and retrieval (RTTHRS) activity thresholds for changing item priority; and the cutoff time limits for update (UDTHRS), retrieval, (RDTHRS), and storage (SDTHRS) items. Observing the statistical summary for data sets one, seven, and nineteen, the two most influential parameters were NITMFL and MFLNTH. Changing MFLNTH from ten to fifty decreased the number of accessions from 2072 to 1196. This is a result of requiring more transactions to transpire prior to attempting reserve record creation thereby preventing unnecessary reorganization. When the MFIL was sufficiently full and the required number of transactions had been performed, the MFIL was replaced regardless of whether the item in the MFIL was in the reserved area. Changing the value of NITMFL from 0.5 in data set seven to 0.9 in data set nineteen, decreased the number of accessions from 1196 to 1028. This is because fewer reserved records were created as the threshold was increased, even though the number of priority changes remained approximately constant. The remaining parameters "fine-tuned" the interface activity by determining the amount of effort allocated to modifying the item priorities. Increasing the value of UPTHRS and RTTHRS reduced I/O activities by forcing fewer priority changes. The I/O accessions eliminated were those required to perform the self-organization of the data set. Increasing the parameters UDTHRS, RDTHRS, and SDTHRS reduced I/O traffic by extending the cutoff date after which an item is checked for item priority change. Once at a certain priority, the item remained in that classification for an extended period. Again the eliminated I/O was at the expense of self-organization.

Attempting to establish these parameter settings require some guidelines. If the data set is dynamic with the frequency of various items varying rapidly over time, low values of UPTHRS, RTTHRS, UDTHRS, RDTHRS, and SDTHRS should be used to allow for quick reaction to the changing item priorities. If the transactions appear to reference the same set of items for an extended time period, then the UDTHRS, RDTHRS, and SDTHRS should be large to allow the items to remain at their respective priority level for longer periods. In addition, if information is known about those item to be requested and those less frequently requested, then the UPTHRS and RTTHRS values should be set to prevent the less frequently accessed items for being considered for reclassification. The setting of the UPTHRS and RTTHRS values should be based on the expected item activity level for those items referenced. The creation of reserved records is dependent on the values assigned to MFLNTH and NITMFL. Whenever an item is placed in the reserved area it is both logically and physically relocated, implying that the item is on a collision chain. Consequently, every time an item is placed in the reserved area a minimum of two accessions is required; one to retrieve the item and one to restore the home record with an updated CRL, assuming the reserved record is in core. Constructing reserved records becomes an expensive process when low MFLNTH and NITMFL value are used. To emphasize, this extra activity is in addition to any activity required to change item priorities.

5. Conclusions

The experimental results indicate that I/O traffic can be reduced by an internal assignment of items to various priority classes. Moreover, if the records containing these items can be kept dense(a function of the relocation routine), I/O traffic can be reduced further. If a memory area is set aside for storing the most active items, substantial I/O savings can be made. Naturally these results hold for data sets having a majority of the activity associated with a subset of that data set. The amount of reduction in I/O is a function of the number of items per record, which in turn is a function of the item length, track space, compression savings, and IIF space required.

The compression results are encouraging in that storage requirements can be reduced. However, compression systems in themselves are nothing new. The use of the compression algorithm with the self-adaptive capability to produce fewer I/O accessions is unique. Moreover, implementing both a self-optimizing program to set parameters for acceptable compression limits and a tree searching program to select a particular algorithm is unique.

Reserved records as used in this set of programs cannot be considered for any future implementation. Instead, a set of telescoping priority classes developing a working set data set should be used. The threshold levels could be set dynamically at execution time to improve system performance.

In summary, for data sets of thousands of items having a substantial activity on a subset of that data set, the application of an interface described in this paper would reduce I/O traffic, storage costs, and user delay time. Cache memory assignments can be easily and efficiently made for further reduction in cost and delay time. Finally, self-adaptive, self-organizing systems indicate performance improvement over strictly manually directed system.

6. References

[1] Allen, R. A., Omnibus: A Large Data Base Management System. AFIPS, FJCC (1968), 157-169.

[2] Amidon, E. L and Akin, G. S. Algorithm Selection of the Best Method for Compressing Map Data Strings, Comm ACM 14, 12 (Dec. 1971), 769-774.

[3] Andrews, C. A. and Schwarz, G. R. Analysis and Simulation of Data Compression Techniques. Proc. Nat. Telemetry Conf. (1967), 57-64.

[4] Andrews, C. A., Davis, J. M., and Schwarz, G. R. Adaptive Data Compression. Proc. IEEE 55, 3 (March 1967), 267-277.

[5] Coffman, E. G. and Eve, J. File Structures Using Hashing Functions. Comm. ACM 13, 7 (July 1970), 427-432.

[6] Denning, P. J. and Schwartz, S. C. Properties of the Working-Set Model. Comm. ACM 15, 3 (March 1972), 191-198.

[7] Dodd, G. C. Elements of Data Management Systems, Computing Surveys 1,2 (June 1969), 117-133.

[8] Frisch, J. Bit Vectors Vitalize Data Retrieval. Data Dynamics (August 1971), 37-42.

[9] Humphrey, A. L. and Munro, W. G. Management Information Retrieval. Computer J 13, 2 (May 1970), 127-130.

[10] Kortman, C. M. Redundancy Reduction - A Practical Method of Data Compression. Proc. IEEE 55, 3 (March 1967), 253-263.

[11] Lefkovitz, D. File Structures for Online Systems, Spartan Books, New York, 1969.

[12] Liu, H. A File Management System for a Large Corporate Information System Data Bank. AFIPS, FJCC (1968), 145-156.

[13] Marron, B. A. and deMaine, P. A. D. Automatic Data Compression. Comm. ACM 10, 11 (November 1967), 711-715.

[14] Price, C. E. Table Lookup Techniques. Computing Surveys 3, 2 (June 1971), 49-65.

[15] Schwartz, E. S. and Kleiboener, A. J. A Language Element for Compression Coding. Information and Control 10, 3 (March 1967), 315-333.

[16] Schwartz, M. H. MIS Planning. Datamation, 16, 10 (Sept. 1970) 28-31.

[17] Weber, D. R. A Symposium on Data Compression. Proc. Nat. Telemetry Conf. (1965), sess. 1, 9-16.

[18] Williams, J. G. Storage Utilization is a Memory Hierarchy When Storage Assignment is Performed by a Hashing Algorithm. COMM. ACM 14, 3 (March 1971), 1972-175.

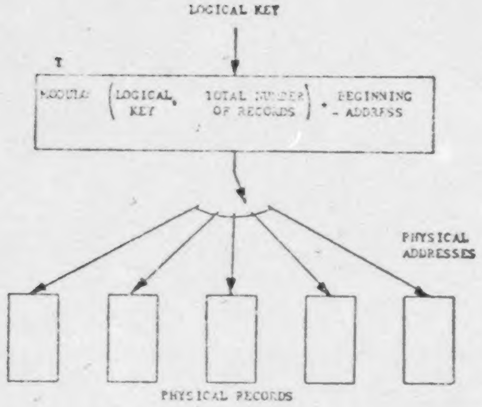

LOGICAL KEY
T
MODULO
LOGICAL KEY
TOTAL NUMBER OF RECORDS
+ BEGINNING ADDRESS
PHYSICAL ADDRESSES
PHYSICAL RECORDS

Figure 1.
Hashing transformation, T, uniformly mapping logical key to physical address.

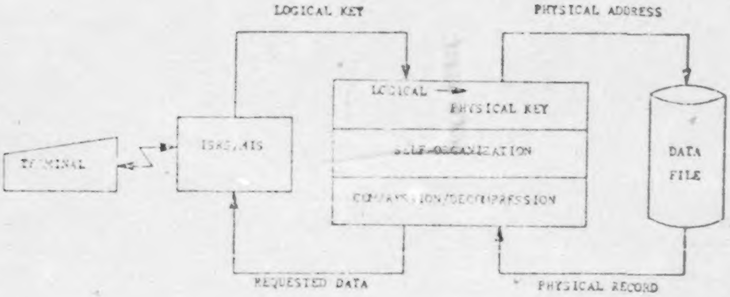

LOGICAL KEY
PHYSICAL ADDRESS
TERMINAL
ISRS/MIS
LOGICAL → PHYSICAL KEY
SELF-ORGANIZATION
COMPRESSION/DECOMPRESSION
DATA FILE
REQUESTED DATA
PHYSICAL RECORD

Figure 2.
ISRS/MIS interface system.

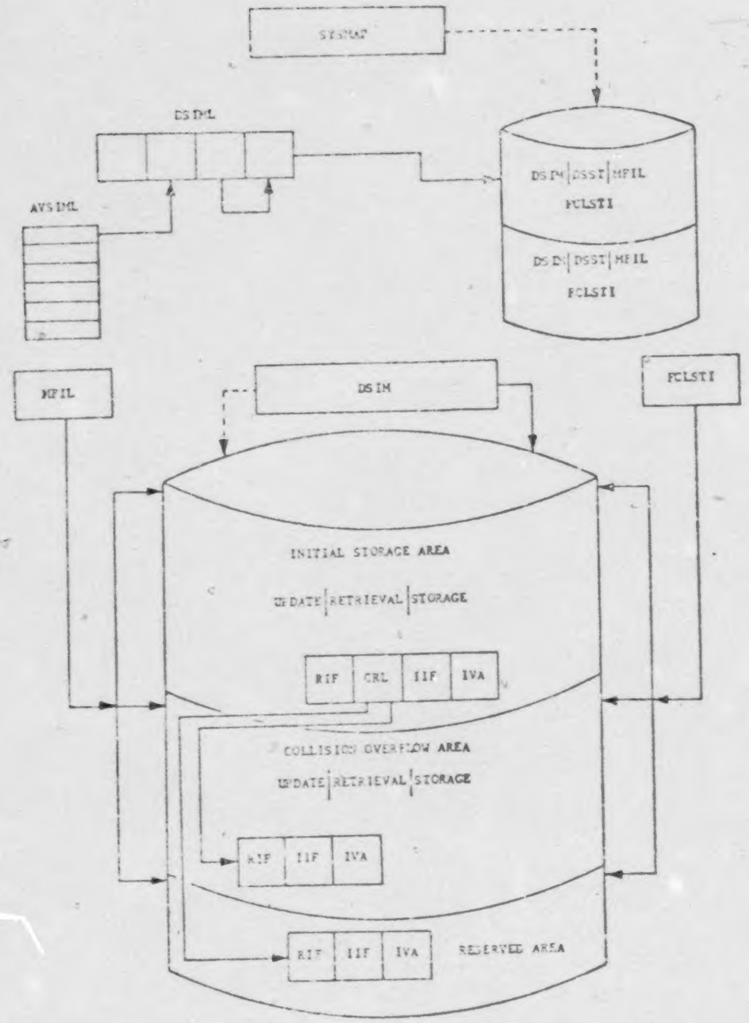

SYSMAP
DSIML
AVSIML
DSIM | DSST | MFIL
FCLSTI
DSIM | DSST | MFIL
FCLSTI
MFIL
DSIM
FCLSTI
INITIAL STORAGE AREA
UPDATE | RETRIEVAL | STORAGE
RIF
CRL
IIF
IVA
COLLISION OVERFLOW AREA
UPDATE | RETRIEVAL | STORAGE
RIF
IIF
IVA
RIF
IIF
IVA
RESERVED AREA

Figure 3.
Logical table relationship.

A FOCUS ON THE ROLE OF THE DATA MANAGER

RUTH M. DAVIS
Director
Institute for Computer Sciences and Technology
National Bureau of Standards, U.S. Department of Commerce
Washington, D.C. 20234

Welcome, ladies and gentlemen. It is a pleasure to welcome you on behalf of the Acting Director of the National Bureau of Standards, and on behalf of my Institute for Computer Sciences and Technology to this Second National Symposium on the Management of Data Elements. I hope that your meeting as a group on this subject will become an annual event here at the Bureau, because your topic--management of data elements--is of extreme importance and plays a key role in the proper use of computers and in the proper handling of information.

Ten years ago, the topic of information science and computers would not have been discussed in any scientific meeting of any repute. This past summer, however, Vice President Rockefeller brought together about 15 scientists to try to identify those major areas of science and technology which the new scientific advisory apparatus in the government should be concerned about and which should be brought to the attention of the President. There were seven major issues discussed including the typical items: food, energy, national security; but there were two new items which impressed me. One was information technology and productivity, and the other was international trade and information and technology transfer. I thought it extremely important to note because that was the first time in such a topical discussion that these items appeared, and I thought that the linkage was very interesting; namely, that information technology was tied to productivity.

I was told by several members of the group that they spent seven and a half hours with the Vice President. That, I think, is an extremely important change from ten years ago and even from five years ago. Si Ramo, one of the leaders of the group, recently told me that he believes that information technology, computer technology, and the handling and use of information are vital to us today, and are key to the major advances that we will be making as a unified society and as a unified nation in the future.

This recent emphasis on information technology is significant but frequently the link between information technology and data base management gets lost. People rarely use the phrase "data base management" when they are discussing information technology. One of the problems is that "data base management" does not connote its importance and its real meaning to those who are dependent upon it. For example, when I have suggested to scientists that data base management systems are vital to good information handling and to the advance of information technology, a non-computer data base management information scientist replied that "DBMS" means "damn bad management support" to him. We have to communicate to scientists our knowledge that data base management systems are the core to some of the problems that now exist and are crucial to some of the successes that we would like to create in the near future.

In addition, the relevance of data base management systems to management must be emphasized. I believe that organizations of any type that will exist in 1985 will have a much higher ratio of computers to men than is typical today, and that successful management of these organizations is going to be dependent upon managers knowing more about computers and how to work with them than they do today. Managers of today take an unbelievable amount of pride in disassociating themselves from their computers and their information systems organizations. They often refuse to admit that their success in management is dependent upon data base management and data base management systems.

Generally, management science or management development texts recommend that a manager have experience in every important aspect of the company before becoming involved in senior management. What we have today is a paradox--managers proudly say "I don't know anything about my computer facility. I leave that up to my data management or my computer facility manager." Those two ideas are contradictory, and one of the reasons that I think management of organizations today is getting worse rather than better.

Managers involved in computer systems should be streaming up the line to senior management positions in companies in the traditional management way, while getting experience in other areas as well. Too often a computer facility manager or a data base manager has reached the end of his career in management and has nowhere to go except to the same kind of job in another organization. Since the management of companies is becoming more and more dependent on computers, data base management systems, and the manipulation of information and data, we as computer specialists, data base managers and information scientists should take the lead in making sure that data base and computer facility managers are part of the road to management in any organization.

By 1985, the rate of change and the pace of work in most organizations are going to be set by computer capability and not by worker incentives and responses, and the management of computer-paced operations is going to be radically different than the management of man-paced operations. We have evidence of this today but have not put the facts together in other than anecdotal form to prove this conjecture. Obviously, the role of data base management and the data managers will be extremely important.

Management trained to make decisions under conditions of uncertainty due to inadequate information is obsolete management. Today management is beginning to make decisions based on the synthesis and association of data that has been manipulated by computers and on the use of data that has undergone some kind of modeling process. Managers of the future will be more dependent upon computer processing of data because they will be faced with more ill-structured rather than well-structured problems.

That conjecture about management in 1985 highlights the very important functions of data control and of understanding the role of data as it is collected. The importance of standardization of data elements, the problems of controlling the flow of data as it becomes information and is manipulated by the system, the problems of data dissemination, data display, and data control when data becomes the output of the system, have become the elements which have given rise to today's "privacy problem." The privacy problem is really based upon inadequate data control and data handling in computer systems and in large manual systems. A mix of legal, economic and technological solutions are being proposed. Legal solutions are needed because the privacy of individuals is involved--something that technology has neither the responsibility nor the capability of resolving.

However, technological solutions are getting inadequate attention. It is absolutely essential for management to work with computer facility managers and the data base managers to determine and isolate the overall problems within an organization that have to be resolved in order for that organization to handle its information properly.

The technical types of solutions that are being proposed, I think, are fairly good ones. One solution, which is pseudo-technical, pseudo-legal and totally desperate is--"for goodness sakes, don't collect data." I have a very strong feeling that one of the primary outcomes of the Privacy Act will be the collection of less data. That is good in some instances, but bad in others.

The second problem that is being examined and resolved is pseudo-technical and pseudo-managerial in its handling of the opposite end. How can you purge data that has been collected in the past when you have not been exercising good control over it when it is collected? The problems of purging data encompass the lack of standardization, the lack of control over data elements, and the lack of documentation and description in qualifiers that are put on the data. For example, when you set up a legal requirement to purge all data relating to criminal records of people five years after those records have been generated, it is pretty impossible to purge that data when there is no data indicating when the criminal record originated.

Another problem relating to the privacy issue is maintaining control over a computer system when there are large data bases and a tremendous flow of information through the computer system. Nothing better has come our way then the traditional process of auditing, originating in the world of CPA's. You hear a lot of talk today but there is very little action on auditing of computer systems. This probably is going to be the key to getting the credibility that we need with respect to data base management, data base control, and even the managing of data elements.

Auditing is a negative word with after-the-fact connotations. Actually there are three kinds of functions we need to deal with, and we have used three words in the Bureau of Standards because we can't find one that works. We have used accountability, auditability and fidelity of computer systems.

Accountability involves assuring the correctness of products and services to gain the respect and the credibility of your public, your constituency and your management. Auditability refers to conducting independent reviews of computer systems that are comparable to those employed by banks and a few other organizations in the country. Fidelity of computer systems has the very important connotation of real time. How do you, on a real time basis, keep probing, sampling and testing your computer system to assure that it is performing its required functions accurately? Secondly, how do you assure that the things that your computer system is not supposed to do with your data are not being done?

These three functions put together--no one has thought of one word to describe them--are really key to all of the kinds of data base management controls that you have been working on for so many years, and that still need to be worked on. We need to talk about them more in order to explain their complexity and importance and the progress that has been made by this community in their handling. The privacy problem, for one, will not be resolved without the creation and implantation of these kinds of controls on data bases and without data base managers taking the lead in describing the necessary controls and then implementing them.

Just as the privacy problem will not be resolved until this kind of data base management function is performed, we will not be able to stop the kinds of problems that are deluging us at the moment and will continue to deluge us. These are the problems of data base management in large systems which affect individuals directly perhaps even more than the privacy problems affect individuals. I refer to those problems that are becoming highly visible in large funds dispersing systems, such as SSI (Supplementary Security Income) and in credit systems where computer fraud, mixed-up management, and inaccuracies affect an individual because wrong financial data is maintained about him. I expect a great deal of public and press attention to be devoted to instances where individuals are harmed because of errors in dispersing funds or in credit ratings.

Another problem area of importance is data accountability and data base management in real time control systems. Most of the data base managers are not involved in real time control systems. That is wrong from my point of view. As real time control systems become more dependent on computers than on manual probing and interaction, there are going to be data base management problems. There will be public safety problems associated with real time air traffic control, rapid transit systems and nuclear power plant operations as more of these systems are utilized.

For example, computers controlling nuclear plant operations on a real time basis will have available much more data, both to control the reactor through efficient fuel utilization and to prevent catastrophic accidents. Public reaction to the possibility of problems in these crucial areas involving real time control will dramatize the need for accountability, auditability and fidelity of computer systems. Is the data correct as it goes through the systems and are controls exercised by data base managers?

The importance of data base managers and data base management systems dictates that we find a way to talk to the unconvinced, the decision makers and those who don't recognize yet the importance of the subject. I consider the proper handling of these systems to be the key to our future in handling our society and our individual problems.

A Proposed Standard Routine
For generating
Proposed Standard Check Characters

Paul-André Desjardins [1]

Hospital Saint-Michel-Archange
Quebec, P.Q., Canada

Some methods of generating check characters have become "de facto" standards. Unfortunately, they have many inefficiencies built-in which only the infancy of computer information processing could excuse. Moreover the lack of a true industry-wide standard could mean a check mate to anyone involved in data interchange. So let us define some new standards and implement them in a single ANSI COBOL routine which generates the desired check character. A full listing of a proposed routine is presented.

Key Words: Check character; check digit; COBOL routine; code; error-detecting; random error; self-checking; standard; transcription error; transposition error.

1. Introduction

A check character is one that is appended to a code as an additional character which serves the purpose of checking the consistency or validity of the code when it is recorded and transferred from one point to another. It is derived by using some mathematical technique (algorithm) involving the characters in the base code. It provides the capability of detecting most clerical or recording errors. These errors are categorized in four types, i.e. transposition errors (1234 recorded as 1243), double transposition errors (1234 recorded as 1432), transcription errors (1234 recorded as 1235) and random errors (1234 recorded as 2243) which are multiple combinations of transposition and transcription errors. [1] The base code with its check character appended is then said to be self-checking.

Unfortunately so many different algorithms have emerged that if two shops wish to exchange self-checking codes they will also have to exchange the precise method used in order that their respective computers be programmed accordingly.

A standard method sure is the solution to the problem. What then is the ultimate algorithm everybody would agree with?

---

[1] Systems Analyst

2. Some popular methods and their weaknesses

2.1. Modulus 10 Method

This method is used for example for the canadian SIN:

basic SIN : 2 1 8 6 2 2 1 3
factors (right to left): 1 2 1 2 1 2 1 2
multiply: 2 2 8 12 2 4 1 6
add the digits: 2+ 2+ 8+ 1+ 2+ 2+ 4+ 1+ 6 = 28
divide by 10: 28/10 = 2 remainder 8
substract from 10: 10-8 = 2 which is the check digit

So the full self-checking SIN is 218622132

2.1.1. Weaknesses (see table 1)

The method cannot detect the double transposition error, for example 218226132 instead of 218622132. Try it. It even fails to detect when 0 is transposed with 9 and conversely i.e 2/90 or 2.2% of all possible simple transposition errors.

2.2. Modulus 11 Method

This method goes like this [3] :

basic code : 9 4 3 4 5 7 8 4 2
factors (right to left) : 4 3 2 7 6 5 4 3 2
multiply and add products : 36+ 12+ 6+ 28+ 30+ 35+ 32+ 12+ 4 = 195
divide by 11 : 195/11 = 17 remainder 8
substract from 11 : 11-8 = 3 which is the check digit

So the full self-checking code is 9434578423

2.2.1. Weaknesses (see table 1)

The remainder of a division by 11 goes from 0 to 10. The special rule is introduced that when the remainder is 0 you do not substract from 11. When you do substract the result is between 1 to 10 and you must eliminate beforehand all the codes generating the check digit 10.

Table 1. Efficiency of the classical methods. Overall efficiency is defined as the sum across all types of errors of efficiency * frequency.

Method ∇	Error type	Transcription	Transposition	Double Transposibion	Random	Overall Efficiency
	Frequency [4]	86%	8%	1%	5%	--
1	Modulus 10 Weights: 1,2,1,2,1,2	100%	97.8%	—	90%	98.3%
2	Modulus 11 Weights: 7,6,5,4,3,2	100%	100%	100%	91%	99.5%

3. Proposed methods

3.1. Modulus 10

Let us take our previous modulo 10 example and see the difference:

basic code : 2 1 8 6 2 2 1 3

factors (right to left): 9 7 3 1 9 7 3 1

multiply and add products : 18+ 7+ 24+ 6+ 18+ 14+ 3+ 3 = 93

divide by 10: 93/10 = 9 remainder 3

So the full self-checking code is 218622133. The efficiency of the method is given in table 3.

Table 3. Efficiency of proposed modulus 10 method.

Error type	Transcription	Transposition	Double Transposition	Random	Overall efficiency
Approx. frequency [4]	86%	8%	1%	5%	---
	100%	88.9%	88.9%	90%	98.5%

3.2. Why is it better?

3.2.1. The base of the system

Whith a modulus 10 system the check-digit is always between 0 and 9 so you don't have to reject any code beforehand or find a complicated scheme [6][7][8]. This sole reason calls for a modulus 10 method if only we can give it an efficiency close to that of the classical modulus 11. Alter all you are never 100% sure that the code with a valid check digit is valid. So we feel that a loss of efficiency in the 1% range is quite permissible.

3.2.2. Take products as they are

Have a look at table 2 on the next page and remember it is a must to catch 100% of transcription errors since they are an overwhelming majority. If we sum up the digits of the products, weights 1,2,4 5,7,8 are the safe ones. If we don't, only weights 1,3,7 and 9 are usable. We choose to take products as they are because it is easier to implement. But summing up the digits of the products is a scheme that deserves more careful examination. Let us only note that it can't detect any transposition between 0 and 9, that factors 1 and 4 do not detect transposition between 0 and 6, that factors 1 and 8 do not detect transposition between 1 and 6, etc.

3.2.3. Transposition errors

Any transposition between digits that have a difference of 5 will be undetected because the differences between any two weights i.e. 2,4,6,8 multiplied by 5 gives 10,20,30,40 which all give 0 modulo 10. There are ten such transpositions (0,5), (1,6), (2,7), (3,8), (4,9) so efficiency is 80/90 = 88.9%.

3.2.4. A real modulus

By definition a modulus is the remainder of a division. We do not substract from 10, a useless operation which is probably a reminiscence of the good old non-electronic days when the self-checking code was verified by giving the check-digit a weighing factor of 1, working back all calculations and looking if the remainder of division by 10 was 0. The proposed method can be implemented without even dividing because the remainder of a division by 10 is always the units position of the dividend.

Table 2. Products between possible weights and digits. Note that modulo 10 all you need to consider is the units position of the products. Underlined are the digits that are recurring linewise so that the weight cannot detect all transcription errors. For example weight 2 does not detect transcription of 5 for 0, 6 for 1, etc.

WEIGHT \ DIGIT	0	1	2	3	4	5	6	7	8	9
1	0	1	2	3	4	5	6	7	8	9
2	0	2	4	6	8	10	12	14	16	18
3	0	3	6	9	12	15	18	21	24	27
4	0	4	8	12	16	20	24	28	32	36
5	0	5	10	15	20	25	30	35	40	45
6	0	6	12	18	24	30	36	42	48	54
7	0	7	14	21	28	35	42	49	56	63
8	0	8	16	24	32	40	48	56	64	72
9	0	9	18	27	36	45	54	63	72	81

Table 2a. Sum of the digits of the products as in the classical modulus 1[illegible] method. If the sum exceeds 9 the digits are summed up again.

WEIGHT \ DIGIT	0	1	2	3	4	5	6	7	8	9
1	0	1	2	3	4	5	6	7	8	9
2	0	2	4	6	8	1	3	5	7	9
3	0	3	6	9	3	6	9	3	6	9
4	0	4	8	3	7	2	6	10= 1	5	9
5	0	5	1	6	2	7	3	8	4	9
6	0	6	3	9	6	3	9	6	12= 3	9
7	0	7	5	3	10= 1	8	6	13= 4	11= 2	9
8	0	8	7	6	5	4	12= 3	11= 2	10= 1	9
9	0	9	9	9	9	9	9	9	9	9

3.3. Modulus 23

This method goes like this:

basic code : 2 1 8 6 2 2 1 3
factors (right to left) : 8 7 6 5 4 3 2 1
multiply and add products : 16+ 7+ 48+ 30+ 8+ 6+ 2+ 3 = 120
divide by 23 : 120/23 = 5 remainder 5

The check character is "F" as found in table 4 so the full self-checking code is 21862213F. Efficiency is as in table 5.

Table 4. Correspondence between remainder and check-character. Choice of letters is based on Gilligan [2] .

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	
A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	T	U	W	X	Y	Z	

Table 5. Efficiency of proposed modulus 23 method.

Error type	Transcription	Transposition	Double Transposition	Random	Overall efficiency
Approx. frequency [4]	86%	8%	1%	5%	----
	100%	100%	100%	95.7%	99.8%

3.4. Why is it good?

3.4.1. The base of the system

This method capitalizes on the fact that method modulo X catches at least (X-1)*100/X% of random errors. For example with modulus 10 method any digit has one chance out of ten being the right check digit for whatever code. So modulus 10 is good at 9*100/10 = 90% , modulus 11 is good at 10*100/11 = 90.9% and modulus 23 is good at 22*100/23 = 95.7% for catching random errors.

3.4.2. Transcription and transposition errors

Since 23 is a prime number the choice of weighing factors is immaterial [5] . We simply choose to keep them at one digit as in our modulo 10 method.

3.5. Future Work

It would be surprising to see method modulus 23 welcomed for checking all numeric fields since it generates a letter. Its usefulness will be for alphanumeric codes but the author has not yet found a scheme that would produce the same check character whether the code is in ASCII or EBCDIC, etc.

Something can also be done for improving random error detection. The problem is to categorize those errors and find an optimum ordering of the weights [4] [9] .

4. Proposed routine

```
001050 IDENTIFICATION DIVISION.
001100 PROGRAM-ID.     MODULO.
001150 AUTHOR.         REJEAN RACINE.
001200 INSTALLATION.   HOPITAL ST-MICHEL-ARCHANGE.
001250 DATE-WRITTEN.   JULY 16, 1975.
001300 REMARKS. STANDARD ROUTINE FOR GENERATING CHECK CHARACTERS
001305      IT IS USED IN A MAIN PROGRAM BY A CALL 'MODULO' USING A,B,C
001310      WHERE A IS PIC 9(2)  AND CONTAINS 10 OR 23
001315            B IS PIC 9(10) AND CONTAINS RIGHT JUSTIFIED THE FIELD
001320                           FOR WHICH WE WANT A CHECK CHARACTER
001325                           EITHER TO ASSIGN OR VERIFY IT
001330            C IS PIC X     AND CONTAINS AN '*' AT TIME OF CALLING
001335                          IT WILL LATER CONTAIN THE CHECK CHARACTER
001340                          IF THE ROUTINE EXECUTES PROPERLY.
001350 ENVIRONMENT DIVISION.
001400 CONFIGURATION SECTION.
001450 SOURCE-COMPUTER. CENTURY-200.
001500 OBJECT-COMPUTER. CENTURY-200.
002000 DATA DIVISION.
002050 WORKING-STORAGE SECTION.
002100 77  COUNTER      PIC 99   VALUE ZEROS.
002120 77  W-REMAIN     PIC 99   VALUE ZEROS.
002150 77  WADD-PROD    PIC 999  VALUE ZEROS.
002170 77  WQUOTIENT    PIC 999  VALUE ZEROS.
002200 01  MODUL-XX.
002250     02  MODU     PIC 9   OCCURS 10 TIMES INDEXED BY IND-MOD.
002350 01  M-PRODUCT.
002400     02 PRODU     PIC 99  OCCURS 10 TIMES INDEXED BY IND-PR.
002420 01  TAB-CURR.
002430     02  FLD-CRR  PIC X(23) VALUE 'ABCDEFGHJKLMNPQRSTUWXYZ'.
002440     02  FIELD-CORR REDEFINES FLD-CRR PIC X OCCURS 23 TIMES.
002440
002450 LINKAGE SECTION.
002500 77  TYPP          PIC 99.
002560 01  CHECKAR PIC X.
002570 01  CHECKAR1 REDEFINES CHECKAR PIC 9.
002570
002600 01  FIELD-DATA.
002650     02  FLD      PIC 9 OCCURS 10 TIMES INDEXED BY IND-FD.

003000 PROCEDURE DIVISION USING TYPP, FIELD-DATA, CHECKAR.
003050 DEBUT.
003060     IF CHECKAR = '*' NEXT SENTENCE
003070         ELSE MOVE '*' TO CHECKAR GO TO OUT-MOD.
003300     IF  TYPP = 10 MOVE '3197319731' TO MODUL-XX
003400         ELSE IF TYPP = 23 MOVE '1987654321' TO MODUL-XX
003410         ELSE GO TO OUT-MOD.
003420     MOVE ZEROS TO WADD-PROD.
003500     MOVE 1 TO COUNTER.
003550     SET IND-MOD, IND-PR, IND-FD TO 1.
003700 A1. MULTIPLY MODU (IND-MOD) BY FLD (IND-FD) GIVING PRODU (IND-PR)
003710     ADD PRODU (IND-PR) TO WADD-PROD.
003750     ADD 1 TO COUNTER.
003800     SET IND-MOD, IND-PR, IND-FD UP BY 1.
003950     IF  COUNTER < 11 GO TO A1.
004400     DIVIDE WADD-PROD BY TYPP GIVING WQUOTIENT REMAINDER W-REMAIN.
004420     IF  TYPP = 10 MOVE W-REMAIN TO CHECKAR1
004440     ELSE ADD 1 TO W-REMAIN
004450          MOVE FIELD-CORR (W-REMAIN) TO CHECKAR.
004470 OUT-MOD.
004550     EXIT PROGRAM.
```

5. Acknowledgements

The author wish to acknowledge the help of his former employer, the Quebec Health Insurance Board under which part of the material was thinked about and written. Thanks also to Hospital Saint-Michel-Archange that welcomed the idea of submitting the paper and to Réjean Racine who programmed the COBOL routine nicely. Thanks to Ministère de la Fonction Publique where we insured that the routine was running well on an IBM machine. And let me not forget Carole Goupil and Diane Tremblay for patiently typing the text.

6. References

[1] Guide for the Development, Implementation and Maintenance of Standards for the Representation of Computer Processed Data Elements, Proceedings of the First Symposium, p.386 (1974).

[2] Gilligan, M.J., Information System Data Coding Guidelines, ibidem, p.167.

[3] Introduction to I.B.M. 360 Direct Access Storage Devices and Organization Methods, C20-1649, p.56.

[4] Beckley, R.F. An optimum system with "modulus 11", THE COMPUTER BULLETIN (Dec. 67). See also March 72, Aug. 72.

[5] Wild W.G. The theory of Modulus N check digit systems, THE COMPUTER BULLETIN (Dec. 68).

[6] Campbell, D.V.A., A modulus 11 check digit system for a given system of codes, THE COMPUTER BULLETIN (Jan. 70).

[7] Reid, C.J., in Letters to the Editor, THE COMPUTER BULLETIN (April 70).

[8] Andrew, A.M. A variant of modulus 11 checking, THE COMPUTER BULLETIN (Aug.70)

[9] Briggs, T., Modulus 11 check digit systems, THE COMPUTER BULLETIN (Aug. 70).

Addendum to
A Proposed Standard Routine
For Generating
Proposed Standard Check Characters

Paul-André Desjardins

Ministère de la Fonction publique
Québec, P.Q., Canada

We would like to add two more references:

[10] Beckley, D.F., Check digit verification, DATA PROCESSING (July-AUG. 66).

[11] Taylor, Alan, Darmstadt System Eliminates Check-Digit Loopholes, COMPUTERWORLD (17 sept. 75).

Based on Taylor [11], it would be fine to revert the order of the weighing factors in our modulo 10 method, that is 9, 7, 3, 1 instead of 1, 3, 7, 9.

78

Methodology For Development of Standard Data Elements Within Multiple Public Agencies

L. D. England, S. L. Eberle,
B. H. Schiff, and A. S. Huffman[1]

Texas State Department of Public Welfare
Texas Department of Health Resources
Texas State Board of Control
Texas State Comptroller of Public Accounts
Austin, Texas

The authors have presented a case study of an extensive standards development project undertaken among nine Texas state agencies during parts of the years 1973 and 1974. It was financed with federal/state matched funds through the auspices of the U. S. Department of Health, Education, and Welfare, the Texas State Department of Public Welfare, and the Texas Governor's Office of Information Services. The degree of transferability of the developed methodology and technology will be examined for potential implementation in other states.

Key words: Data element dictionary; data element standardization, data standards in public agencies; manual of data elements; standardization methods; state data elements and representations.

1. Introduction

At the first National Symposium on the Management of Data Elements in Information Processing, we presented the technical aspects of using a sophisticated data base management system in support of analyses pursuant to development of standard data elements among several Texas state agencies. In that presentation we described an analytical approach used both to achieve consensus and to gain acceptance of a set of standard data elements and codes.

Today I would like to report the success of that approach and to enumerate some of the factors which we believe were significant in reaching a successful conclusion. Additionally, I would like to discuss some salient problems and to indicate some facets of project transferability.

---

[1] Assistant Chief, Data Systems Bureau,
Systems Analyst,
Systems Analyst,
Programmer Analyst.

2. Organizational Setting

Early in 1973 the Texas Governor's Office of Information Services funded a group of computer and information specialists for the purpose of increasing the effectiveness of computer technology in Texas State Government. This entity, the Office of Information Services (OIS), provided an organizational base and an appropriate core staff of system analysts and programmers. Additionally, concurrent projects helped justify, procure, and implement the use of advanced technology such as data base management systems.

Under the auspices of the Governor's Office, OIS analysts were able to obtain agency attention and cooperation for implementing various "inter-agency" projects. From time to time inter-agency committees were utilized for consideration of matters of common interest to several agencies. Hence, the inter-agency committee was a familiar mechanism in Texas State Government prior to its use for the selection of standard data elements and representations.

3. Personnel Experience

While prior standards experience was not required of staff personnel, relevant experiences of key participants undoubtedly contributed to the successful accomplishment of project goals. One member of the team (England) had benefited from association with Grace Murray Hopper and the Navy's efforts at standardization. Another staff member (Eberle) had done extensive work in analyzing data elements and forms used by Texas state agencies. These experiences, coupled with the computer technology expertise of Schiff and Huffman, eventually led to creation of a support data base management system, which greatly facilitated the analytical approach. Additional knowledge was also gained from previous standards experience in Texas State Government. Already within the Governor's Office a partial standards effort was underway in the criminal justice area; liaison was established early with personnel of this project. Also, several Texas agencies had previously attempted to formulate a set of data element standards, called TIPS (Texas Information Processing Standards). Based on his experiences in HEW, a deputy commissioner[1] of one of the major participating agencies had counseled consideration of a standards effort at the state level. Valuable knowledge and advice were obtained from the data processing manager[2] of another major agency, an individual who had been intimately involved in the process of establishing national data standards for the Manpower Administration. Several trips to Washington, D.C. provided valuable insight into the federal standards efforts, and the advice of NBS personnel was most valuable.

4. Funding

Within our organizational structure, dedicated funding was essential for establishment of the standards project. Monies became available through use of state general revenue funds for matching with federal administrative (so called 50-50) funds. Considerable resources in the form of time and marketing are required to procure such funds; at least this was true in our case. Sufficient funding was obtained to support three analysts, one programmer, and one research assistant; this was necessary in order to accomplish meaningful results in a short period of time.

---

[1] Robert Nakamoto, Deputy Commissioner for Planning and Management Systems, Texas State Department of Public Welfare.

[2] Sam Montgomery, ADP Director, Texas Employment Commission.

5. Organizational Characteristics

Several other factors further contributed to the success of this project. These resulted from an organizational emphasis which paralleled project goals and objectives. Here I would include: 1) a reasonable degree of organizational flexibility so that appropriate liaison could be readily effected, 2) an internal impetus to "publish" or produce usable products, which exactly paralleled our goals, and 3) managerial support for utilizing state-of-the-art technology in solving "real world" problems. Additionally, the project helped provide a favorable image for the sponsoring organizations through its use of an inter-agency committee for the selection of standards, as these standards would affect all participants.

6. Functional Area Concept

A definite plus element was the existence of an inter-agency committee to coordinate state activities in our functional area: health and human resources. This committee, the Inter-agency Health and Human Resources Council, commissioned the establishment of a standards task force under its auspices. A functional approach to standards development was favored by project staff. It was believed and later proven that the commonality of interests and problems among participating agencies would aid in the standardization process. It was anticipated that standards development might logically proceed through several functional areas, culminating with a subset of "state-wide" standards - the counterpart of Federal General Standards[1]. The concept is illustrated in Figure 1.

7. Project Progress

The project progressed through tasks planned with usual PERT and GANTT charts, finally publishing a standards manual during the summer of 1974. Nine agencies from the health and human resources area participated; included were all agencies having in-house computer facilities. About eight inter-agency committee meetings were held during the most intense work period. The staff stressed analytic expertise from the start and gradually convinced the committee of the thoroughness of the analyses. All standards were accepted by unanimous vote, and the successful vote rate rapidly increased until some 38 data elements encompassing 215 different agency representations were negotiated. The extent of the analyses is illustrated in Figure 2 which is a partial list of the sets of standards uncovered during the course of the project.

8. Results

The project produced several products for use by participating agencies and others. These included a data element standards manual (see Figure 3 for sample page); a data element dictionary, (see Figure 4 for sample page); and a viable automated data base from which several special reports were generated for the agencies. In addition, I think we demonstrated the effectiveness of the methodology in a multi-agency environment. Subsequently, some of the project staff members have moved into agencies which participated in the standards effort. There they have helped to stress the need for use of the standards. Valuable federal and national liaison was established and the group actually hosted a meeting of ANS X3L8 during February, 1975. Attendant favorable publicity within Texas State Government has been evident.

---

[1] U.S. Department of Commerce, National Bureau of Standards, Federal Information Processing Standards Index (FIPS Pub. 12-2). Washington, D.C. U.S. Government Printing Office. December 1, 1974. p. 37.

9. In Retrospect

Standardization of data elements and data representations is a continuing process. With the phasing out of the Governor's Office of Information Services (a political decision), data standardization on an inter-agency basis in Texas State Government has slowed to a snail's pace. The Governor's Office provided the catalyst and the neutral ground for inter-agency cooperation. Attempts are being made to form an ad hoc committee concerned with data element standardization.

Funding, while always necessary, is required in decreasing amounts once the standardization process is in place. After the initial collection of data for the standards data base, organizational planning, and computer programming, personnel requirements can be reduced. However, due to the methodology of having a "permanent" staff to analyze and present data elements for standardization, a higher funding requirement exists, as compared to other methodologies.

As previously indicated, the two salient factors which added most to the success of the project were the use of a computerized data base manager and a "permanent" staff to provide analysis for data standardization. The use of a computerized data base manager greatly aided the organization of the voluminous data collected for the standards project. It also simplified programming of reports that showed the present use of each data element within the participating agencies. The "permanent" staff provided most of the leg work and analysis. This meant that the costly and valuable time of the participating agencies' data processing personnel was held to a minimum, greatly increasing the acceptability of such a data standards project to the participating agencies.

10. Transferability

I am convinced that the concepts of this project can be transferred to certain other multi-agency environments. The concepts have been described here today and in other available publications. Documentation of the project is available, but needs condensation. It includes: PERT charts, inter-agency committee correspondence, project notebooks, data base specifications, computer program documentation, copies of major reports, and copies of the standards manual. The authors will attempt to meet any requests but are currently severely limited by a lack of resources.

LEVELS FOR STANDARDS DEVELOPMENT

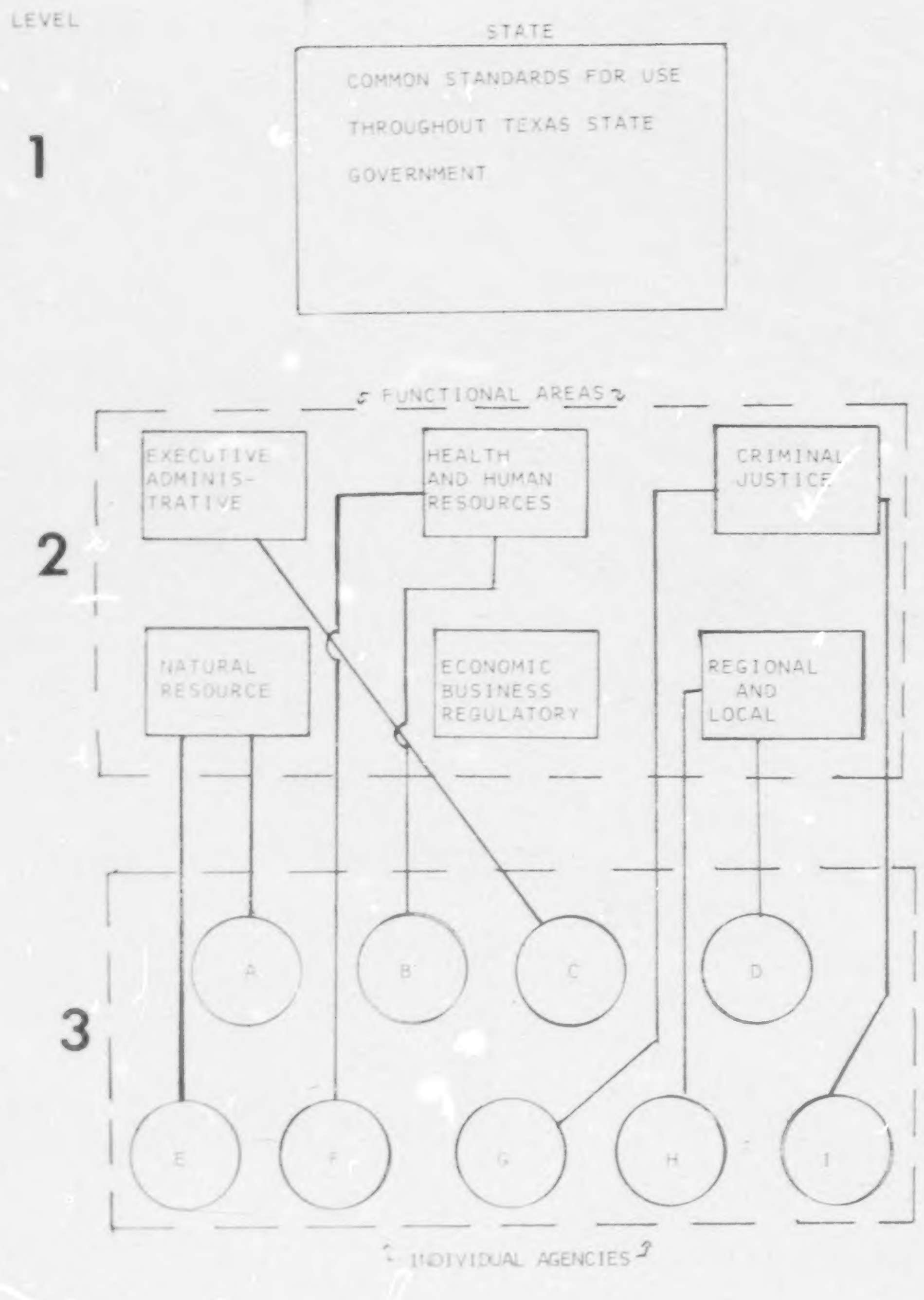

LEVEL
STATE
COMMON STANDARDS FOR USE THROUGHOUT TEXAS STATE GOVERNMENT
1
FUNCTIONAL AREAS
2
EXECUTIVE ADMINISTRATIVE
HEALTH AND HUMAN RESOURCES
CRIMINAL JUSTICE
NATURAL RESOURCE
ECONOMIC BUSINESS REGULATORY
REGIONAL AND LOCAL
3
A
B
C
D
E
F
G
H
I
INDIVIDUAL AGENCIES

FIGURE 1

England/Eberle/
Schiff/Huffman

```
LNP 1 59
     1:USOE      U.S. OFFICE OF EDUCATION
     2:RSA       REHABILITATION SERVICES ADMINISTRATION
     3:BNDX      BENEFICIARY DATA EXCHANGE
     4:COMM      COMMERCE DEPARTMENT (1970 CENSUS BUREAU)
     5:CB        STATE COMMISSION FOR THE BLIND
     6:COO       COORDINATING BOARD, TEXAS COLLEGE AND UNIVERSITY SYSTEM
     7:DPW       STATE DEPARTMENT OF PUBLIC WELFARE
     8:IAB       TEXAS INDUSTRIAL ACCIDENT BOARD
     9:MHMR      TEXAS DEPARTMENT OF MENTAL HEALTH AND MENTAL RETARDATION
    10:TEA       TEXAS EDUCATION AGENCY
    11:TEX       TEXAS EMPLOYMENT COMMISSION
    12:TRC       TEXAS REHABILITATION COMMISSION
    13:TSDH      TEXAS STATE DEPARTMENT OF HEALTH
    14:SRS       DATA STANDARDS CATALOG, SOCIAL AND REHABILITATION SERVICE
    15:FIPS      FEDERAL INFORMATION PROCESSING STANDARDS
    16:DOT       DICTIONARY OF OCCUPATIONAL TITLES
    17:MAIP      MANPOWER ADMINISTRATION INFORMATION PROCESSING STANDARDS
    18:RSA       REHABILITATION SERVICES ADMINISTRATION
    19:SICM      STANDARD INDUSTRIAL CLASSIFICATION MANUAL
    20:TIPS      TEXAS INFORMATION PROCESSING STANDARDS
    21:CPA       COMPTROLLER OF PUBLIC ACCOUNTS
    22:CRS       CLINICAL RECORDS SYSTEM
    23:APA       AMERICAN PSYCHIATRIC ASSOCIATION
    24:ASMD      AMERICAN ASSOCIATION ON MENTAL DEFICIENCIES
    25:ICDA      INTERNATIONAL CLASSIFICATION OF DISEASES
    26:HGIS      HIGHER EDUCATION GENERAL INFORMATION SURVEY
    27:SSI       SUPPLEMENTAL SECURITY INCOME
    28:SCM       STATE CONVERSION PROCEDURES MANUAL
    29:SCAR      STATE COMPUTERIZED ACCIDENT REPORTING SYSTEM
    30:ANSI      AMERICAN NATIONAL STANDARD
    31:URS       UNIFORM REPORTING SYSTEM
    32:IOSD      INTERNATIONAL ORGANIZATION STANDARDIZATION DRAFT
    33:COMP      COMPTROLLER AGENCY CODE
    34:TCJS      TEXAS CRIMINAL JUSTICE INFORMATION SYSTEM
    35:DOL       U.S. DEPARTMENT OF LABOR
    36:HEW       U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
    37:DOD       U.S. DEPARTMENT OF DEFENSE
    38:COM       U.S. COMMERCE DEPARTMENT
    39:HUD       U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT
    40:USDA      U.S. DEPARTMENT OF AGRICULTURE
    41:EEOC      EQUAL EMPLOYMENT OPPORTUNITY COMMISSION
    42:PRES      EXECUTIVE OFFICE OF THE PRESIDENT
    43:GSA       GENERAL SERVICES ADMINISTRATION
    44:GAO       GENERAL ACCOUNTING OFFICE
    45:ISO       INTERNATIONAL STANDARDS ORGANIZATION
    46:TBAA      TEXAS STATE AGENCY BUSINESS ADMINISTRATOR'S ASSOCIATION
    47:OIS       OFFICE OF INFORMATION SERVICES
    48:COUN      HEALTH DEPARTMENT COUNTY CODES
    49:GEOG      GEOGRAPHICAL LOCATION CODES
    50:PLCE      GSA PLACE CODES
    51:HBV       HANDBOOK V, USOE
    52:DFIP      DRAFT FIPS
    53:RAMC      RAND McNALLY AND CO. CODE LIST
    54:ANS2      ANSI-DRAFT-NAME FORMATTING FOR INDIVIDUALS
    55:ANS3      ANSI-DRAFT-STANDARD IDENTIFIER FOR INDIVIDUALS
    56:ANA       AMERICAN NATIONAL STANDARD INSTITUTE
    57:IBM       INTERNATIONAL BUSINESS MACHINES CORPORATION
    58:NCOD      IBM NUMERICAL CODE FOR STATES, COUNTIES, AND CITIES
    59:CIUN      CHARACTERISTICS OF THE INSURED UNEMPLOYED
```

FIGURE 2

DATE (CALENDAR & ORDINAL)

TEXAS INTERAGENCY
HEALTH AND HUMAN RESOURCES COUNCIL
INFORMATION SYSTEMS DATA STANDARDS

Page 1 of 1

SEP 1, 1974

IAHHRC STANDARD	X
FEDERAL STANDARD	X
PROPOSED IAHHRC STANDARD	

ELEMENT DEFINITION

1. Calendar date consists of the year, month and day.
2. Ordinal date consists of year and day of year.

CODE STRUCTURE

See Remarks

REMARKS

1. Gregorian (calendar), 6-numeric-filled with two digits each for year, month, and day in that order.
2. Julian (ordinal), 5-numeric-filled with two digits for the year and three digits for the ordinal day.
3. Length of field may be modified if the entire date is not required. See documents referenced in the code explanation.

CODE EXPLANATION

American National Standards Institute	(ANS X 3.30)
Federal Information Processing Standards	(FIPS-4)

FIGURE 3

```
1290 APPLICANT/RECIPIENT RACE
               A CODE OR NAME INDICATING RACE OF APPLICANT/RECIPIENT.
    SEE ELEMENT(S):                           NUMBER OF OCCURRENCES  81
    AGENCY USERS: CB        COO       DOL       HEW       MHMR      OIS       TE
                  TEC       TRC       TSDH      USOE      DPW

0670 APPLICANT/RECIPIENT REFERRED FROM
               NAME OF PREVIOUS AGENCY, INSTITUTION,
               OR PROFESSIONAL (E.G., A PHYSICIAN) WHICH REFERRED
               APPLICANT/RECIPIENT TO PRESENT AGENCY. ALSO USED WHEN
               ONLY CATEGORY OF REFERRAL SOURCE IS GIVEN, E.G.,
               PRIVATE PHYSICIAN, GENERAL HOSP.,ETC.
    SEE ELEMENT(S):  671    690               NUMBER OF OCCURRENCES  21
    AGENCY USERS: CB        MHMR      TRC       TSDH      DPW

0690 APPLICANT/RECIPIENT REFERRED TO ANOTHER AGENCY
               IF APPLICANT/RECIPIENT IS REFERRED TO ANOT
               AGENCY FOR SERVICES, INDICATES WHAT AGEN
    SEE ELEMENT(S):  670    671               NUMBER OF OCCURRENCES   7
    AGENCY USERS: CB        MHMR      TRC       DPW

2359 APPLICANT/RECIPIENT RELATIVES PHONE NUMBER
               TELEPHONE NUMBER WHERE SPECIFIED RELATIVES OF AN
               APPLICANT/RECIPIENT MAY BE CONTACTED.
    SEE ELEMENT(S):                           NUMBER OF OCCURRENCES   2
    AGENCY USERS: DPW

1270 APPLICANT/RECIPIENT SEX
               CODE OR NAME SPECIFYING SEX OF APPLICANT/RECIPIENT.
    SEE ELEMENT(S):                           NUMBER OF OCCURRENCES  82
    AGENCY USERS: CB        COO       DOL       HEW       IAB       MHMR      OI
                  TBAA      TEA       TEC       TRC       TSDH      USOE      DF

0510 APPLICANT/RECIPIENT SIGNATURE
               SIGNATURE OF APPLICANT/RECIPIENT.
    SEE ELEMENT(S):                           NUMBER OF OCCURRENCES  46
    AGENCY USERS: CB        IAB       TRC       DPW

1100 APPLICANT/RECIPIENT SOCIAL SECURITY NUMBER
               THE NINE-DIGIT ACCOUNT NUMBER ASSIGNED BY THE SOCIAL
               SECURITY ADMINISTRATION.
    SEE ELEMENT(S): 1530                      NUMBER OF OCCURRENCES  77
    AGENCY USERS: ANA       CB        DOL       HEW       IAB       MHMR      O
                  TBAA      TEA       TEC       TRC       TSDH      USOE      D

0665 APPLICANT/RECIPIENT TRANSFER?
               USED WHEN QUESTION IS ASKED IF APPLICANT/RECIPIENT
               HAS BEEN TRANSFERRED FROM ANOTHER FACILITY, OFFICE,
               PROGRAM, ETC. WITHIN THE SAME AGENCY.
```

FIGURE 4

The Role of the Internal Auditor in Data Management

Richard H. Fahnline

Office of Management Analysis and Audits
U.S. Civil Service Commission
Washington, D.C. 20037

Internal auditing is expanding to include audit of data processing. They provide an independent and objective look at operations. They can help provide management controls and standards, by participating in ADP planning. For data elements, auditors will want to learn: description, needed and achieved accuracy, where do they come from, what do they cost, what is their value, are there alternatives, who has access, how long are they useful, and what other uses do they have? Systems should document these points for all data elements. Some systems are relatively common to many organizations, like payroll. We need standards for such systems, in part to reduce the workload of developing standards for data elements. Some progress has been made in data standards, but much remains to be done.

Key words: Data management; internal audit; management controls; standard systems.

1. Introduction

Recent years have seen rapid expansion of the role of internal auditors from almost exclusive concentration on financial matters to evaluation of any and all processes. Even more recently, internal auditors have realized they must penetrate the cloak surrounding data processing in all of it's manifestations and audit not only the ostensible results of automated systems, but the systems themselves and the hardware and software supporting them. Many internal auditors find this new, challenging, and perhaps even frightening; and data processing personnel have an even greater adjustment to make. At the same time systems and operations personnel find themselves scrambling to provide adequate security for data and systems, they suffer the instrusions of auditors--a seeming contradiction.

I will review the purpose of internal auditing, as it applies to data processing and particularly in the context of data management. Of course, data management fits into the nexus of data processing, just as data processing subsumes the larger context of the molecular organization--by this I mean an entire government agency, business enterprise, or the like--in all of its simplicity or complexity, and in turn the molecular organization interacts and interconnects with larger and larger wholes. While we might trace relationships from the very specific to the very general, as we approach the latter the influences become ever more tenuous and eventually exceed our present knowledges and abilities; and certainly exceed our present focus of interest. Let us therefore state as an underlying assumption that managers, data processors, and internal auditors must accept responsibility for the general consequences of their specific acts and decisions; that to the extent they can with reasonable effort and care predict them they should do so; and start from there.

2. Internal Auditing

What exactly is an "internal auditor?" Generically, they are auditors subject to the authority of the management they report to. Thus, if a management has a staff of auditors charged with review of operations under that same management, these are internal auditors. If the same management hires or contracts with an outside organization to come in for specific audit purposes, these people are not internal auditors. For example, to the Civil Service Commission I am an internal auditor because I am on the CSC payroll, I audit CSC, and I report audit findings to CSC management. Similarly, our ADP staff has auditors concerned solely with ADP matters: these are internal to the ADP organization. Internal auditors are a resource of the managements to whom they report.

Internal and external auditors define their work in terms of two words: "independent" and "objective." People involved in work identify with it; management needs to look at their work with relatively unbiased eyes, and these are the eyes of the auditor. This does not imply distrust, although auditors should find any breaches of trust. For the same reason managers must use check digits, redundancy, verification procedures, and batch totals, they use auditors to ensure the adequacy and effectiveness of these and other controls.

The mystic of data processing has removed it from many of the management controls applied to manual processing. Specifically, automated systems have avoided or evaded controls over development time, cost limitations, and thorough documentation, to name just a few. They lack performance standards. Most amazingly, they lack codification, so that again and again organizations start from scratch to develop systems which do essentially what tens or hundreds or thousands of other systems already do. Given the precision necessary for data processing, that data processors would fail to apply to their own generic trade the same rigor they insist on from the functions they serve would be ludicrous if it weren't so detrimental both to the data processing organizations and the processed functions.

3. Internal Auditors in ADP

Many, if not most, data processing professionals know of these deficiencies: the literature abounds with documentation and exhortations for change. Almost all ADP managers worth their salaries can explain why they fall so short in these areas: prominent among the explanations is the overwhelming press of work--they are so busy working they don't have time to manage. This is the classic managerial dilemma, not invented by data processing personnel, and allows of only one response . . . the manager must take time to manage. Every other of organizational endeavor has had to accept this; the wonder is that ADP still gets away with this excuse.

Part of the problem is lack of benchmarks. Higher level managers don't know when ADP costs become excessive, when run times fall outside acceptable tolerances, when programming times exceed normal for competent programmers, when testing and debugging takes too long. The processes have been made unfamiliar to higher managers, but unfamiliarity no longer can be accepted. The profession of data processing must develop standards against which it will measure itself, and be measured. This will give managements the tools to control and manage data processing, without necessarily learning to be data processors. And internal auditors will be there, working with the data processors and managers, developing and validating the standards and applying them to day-to-day operations.

This means that internal auditors must participate in the planning for ADP services. They must work with the specification, review, selection, installation, operation, and evaluation of hardware and software, in the sense that they have an input in defining standards and controls. Ideally, an auditor could rely on review of the adequacy and effectiveness of controls to audit an ADP installation. This not only makes the audit job easier, but also reduces the disruption caused by the audit effort and identifies a management on top of its job. The better the management of an operation, the more it uses standard controls for its own purposes; and these also serve the purposes of the auditors. The same applies to applications: auditors consider an entire function, not just the part handled by computer; and the controls and standards established provide a measure of the application's quality.

Because many auditors understand data processing auditing so little, and because data processing personnel must learn the discipline of standards and control, they must all work together to arrive at the mutual goal.

4. Auditing Data Management

In data management the auditor must be concerned with all aspects. Every paper presented at this Symposium will have relevance for auditing. Check characters? The auditor should test whether the characters do what they are supposed to, as they are supposed to. Standard codes? The auditor will review applications to determine whether they use relevant standards, and the validity of exceptions to standards. Each person can apply this approach to each of the presentations: what significance does this have for audit, and does audit have for this?

We can make a very preliminary list of audit points related to data management. These are points an auditor should look for in the documentation, operation, and products of an application, for each data element.

A. Description. Is each data element (field) described unambiguously and clearly for easy reference? Does the description include full delineation of all appropriate characteristics: permissable values, type of data (alpha, numeric, alphameric), exceptions, meaning of values, references?

B. Accuracy. Does the documentation state the expected, necessary, allowable accuracy level of the data? Where inaccuracy occurs? Impact of inaccuracy?

C. Source. Is the source of each data element defined, including the chain of sources where relevant, to promote understanding of the element as well as nomenclature?

D. Cost. What does it cost to collect, convert, use the data? What is the replacement cost? What does it cost to keep the data secure?

E. Value. What is the value--benefit--of the data, or alternatively the cost of not having it?

F. Alternatives. What alternative data could be used, at what cost and saving?

G. Access. Who should have access to the data, and what safeguards should be established? This includes consideration of freedom of information and protection of privacy, plus analysis of the costs--the cost of protecting/releasing information may tip the balance making the data uneconomical.

H. Life. How long is the data useful, how soon is it available in terms of system outputs, and how does this affect the value/cost of the data?

I. Additional Uses. What are the possibilities and economies of systems sharing the data? What savings in data collection, conversion, storage does this achieve, and what costs/complexities does it introduce?

We can add to this list. As it stands, however, it illustrates the interactions that systems designers and auditors should consider. For example, one data element may fit the needs of no one system perfectly, but represent the best balance among a group of systems. Another data element may be too short-lived for the dynamics of the system it serves. Acceptable accuracy may impose unacceptable costs.

These audit points raise a question of their own: determining these points for each data element, weighing and documenting them, imposes a cost which must weigh in the balance. Superficially, it would seem that by skipping careful element-by-element justification we would save considerable time and cost, thus reducing overall expenses and making better use of our scarce resources. Indeed, we could carry data element justification too far and defeat our purpose. But we must analyze the data elements we use, and to an extent far greater than we have done before. We must formalize the selection of data elements to include how we select them, who selects them and authorizes changes, and how often we repeat the process to ensure the continued validity of prior analyses.

5. Aids

Fortunately, we can capitalize on several advantages, or aids, inherent in the situation. First, most systems undergo a series of modifications and changes: we can use an evolutionary approach to achieving adequate data control and justification. This permits us to transition while keeping our present systems whatever their defects until we can modify or replace them. Similarly, performing the analysis for one system benefits other systems to the extent that they share data or have the same data elements.

This brings us to the second advantage, but one we must create. Many systems in different organizations duplicate each other. In the trite but tragically true phrase, we keep reinventing the wheel in our data systems. For example, almost every organization has a payroll, but we don't have standards for what data to include in payroll systems. We can say with justice that this condition existed in the pre-computer days; but why does this lack continue today? The age of computers makes possible in a way never bever before true the design of a system once with use in an unlimited number of organizations. But what do we have? Unlimited numbers of systems. We don't even know approximately how much per year per employee a payroll system should cost. Even if we accept that not all payrolls can be the same, we must admit that a relatively small number of such systems would take care of the vast majority of organizations. The same applies to many types of applications: personnel, inventory, accounts receivable, and so on and on.

A number of organizations use the same payroll system, some because they buy the service from a service bureau, or because they modified their system from the same ancestral system. Still, we don't know what constitutes a good payroll system. Only in those circumstances where service bureaus offer the same or similar services can we even compare cost. What we need is research into evaluation standards, not only for data processing systems in general but also for specific types of applications such as payrolls. Even if these provide a range of standards, we will have a better basis for evaluating our individual systems.

The second advantage, then, is that we can cooperate in setting standards for what data to use, and how to use them. We do not have to duplicate in organization after organization the effort and cost of developing methods for setting standards (standards for standard-setting, if you will), much less developing the individual standards themselves. We have a wealth of different systems in the various application areas to compare and find the true virtues and defects of each, in order to identify and use the best of each. By working together, organizations can share the cost and do what probably only the largest, if even they, could do for themselves.

Our emphasis here today is on data--but the job must include standards for data elements and data collection and data conversion and all phases of data processing, for the applications and the technology. We need standards for systems design, for programming, for operations. We need standards for all levels of managers to manage with.

6. Status

The data processing community has already made modest progress with data standards. We have some standard data elements, although we don't have standards for use of them in specific types of systems. We don't have standards for evaluating applications. We need such standards, and we need them yesterday.

It is entirely appropriate that we should be meeting under the auspices of the National Bureau of Standards and the American National Standards Institute. From here on, the primary word, the only word, must be "standards."

It's not just that the internal auditors want standards, that they have this need to justify their jobs and so they impose restrictions on organizations and people and systems. The truth is, we gather into organizations to accomplish objectives. The better we define those objectives and measure progress toward them, the better we do our jobs. Call it management by objectives, call it intelligent data processing, call it what you will, it's a goal toward which we all must strive. We have a long way to go. The internal auditor can help you get there.

SEMANTIC CODING AND DATA ELEMENT CHARACTERIZATION IN MEDICAL COMPUTING

Elemer R. Gabrieli, M.D., F.C.A.P.[1]

Clinical Information Center
E. J. Meyer Memorial Hospital
Buffalo, New York 14215

Medicine is concerned about the growing information crisis in clinical practice. The main symptoms are declining quality of medical decisions and fragmented patient care. Although information science has clarified the related concepts and the technology is available to resolve this crisis, the interface between medicine and the technology is still not sufficiently developed. This interface requires a comprehensive, well-constructed Medical Lexicon in which each lexical entry is purposefully coded. A multidisciplinary national task force should be created, authorized, and funded, to develop, code, and maintain a Medical Lexicon. The second interface task is to develop criteria for a Medical Data Element Dictionary, which adequately characterizes all the data elements. In order to conserve the humanitarian and social values, medical information systems should be reviewed and accredited. Only data centers with a sufficiently documented Medical Data Element Dictionary should be certified to continue handling confidential medical data. A highly visible coordinating leadership in the man-machine interface area may be necessary to end the distressing medical information crisis.

Key words: artificial intelligence; coding; cognitive memory; confidentiality; data characterization; man-machine interface; medical coding; medical computing; medical information systems; medical lexicon; medical taxonomy; medical terminology; privacy; right-to-privacy; standardized medical nomenclature.

---

[1] Chairman, Joint Task Group on Confidentiality of Computerized Medical Records.

1. Information crisis in clinical medicine

Only two generations ago, a well-trained physician mastered all aspects of medicine. One generation ago, the principles of diagnosis and therapy gathered during the medical school years remained the high standards for bedside practice throughout the life of the practitioner. Today, owing to the recent spectacular advances in biomedicine, the useful life of medical knowledge has been greatly reduced. The present average half-life of medical knowledge is only about five years [1]. Consequently, the quality of diagnostic/therapeutic decisions is declining. There is a growing gap between the potentially available and the actually used scientic information. Today's clinical medicine is in an information crisis. It is now apparent that the traditional flow of new information from the researcher to the bedside decision maker is no longer adequate. The practitioner can no longer cope with the pace of progress.

Another alarming symptom of the crisis is the fragmentation of patient care. Specialization has resulted in specialty-centered patient management, often at the expense of rational care plans or coordinated treatment. Communication across the borders of the specialties has been eroded. Longitudinal follow-up care by the same family physician has been largely replaced by specialists' care which is concerned only with the current clinical problem. This growing fragmentation of patient care prompted the recent renaissance of family medicine with the hope that such primary care physicians will improve care coordination and revive longitudinal thinking.

The obvious solution of the information crisis is an effective use of modern information sciences and technology. Physicians should no longer be expected to practice entirely from memory. An interactive medical information system should support the diagnostic decisions and the therapeutic choices. Shared clinical experience should be readily available so that the practitioner can make the "best" decisions [2]. Information retrieved should be a part of rational decisions. The role of the Physician should be high level professional judgement instead of memory-limited recall effort. Medical Information systems are badly in need. Perhaps they are long overdue.

2. Man-machine interface problems

Medical communication is in natural language. Diagnoses, conditions, signs and symptoms are expressed in natural language, with both the flexibility, and the ambiguity of human communication. The accurate transformation of natural language terms into machine-compatible semantic equivalents is a real challenge of today's medicine. Conceptually, this transformation requires a formal Medical Lexicon which would contain all the valid medical terms used in clinical communication. Such a Medical Lexicon would have to have one (and only one) entry for every significant medical datum. The Lexicon should cover all fields of clinical medicine and biomedical research. Each entry would represent one clinical idea, concept, expression, or term. Actually, this Medical Lexicon would represent the complete potential data base of biomedicine. To each lexical entry, a practical, unique code value would be assigned. This code value would be the semantic equivalent of the natural language term within the information system. It has been proposed that the process of encoding, per se, should be automated [3]. Manual coding is notoriously inaccurate, often arbitrary. Automated encoding, i.e., computer-based retrieval of the proper code, should greatly reduce the human error of coding.

A Medical Lexicon would have to provide the necessary flexibility at the human end, and a rigid consistency at the machine's end. The former is imperative for human acceptance of the machine, while the latter is critical for semantic retrieval of shared clinical experience [2].

3. Available medical vocabularies

The most comprehensive, and perhaps the most current listing of diseases, signs, and symptoms, conditions, injuries, and medical procedures is the Eighth Revision International Classification of Diseases Adapted for Use in the United States (ICDA) [4] and its hospital-oriented variant, the H-ICDA-2 [5]. However, ICDA was not designed for clinical use. It was intended to serve as a statistical tool. The scheme of ICDA is the reorganization of all medical terms into 1050 categories. There are at least 50,000 semantically distinct, valid medical terms used in clinical communication. It was not the concern of the architects of ICDA to provide every term with a different code. Quite to the contrary, their aim was classification of diseases, so that ICDA coding can place a patient into a set of more or less similar cases. This classifying approach is somewhat contrary to the medical goal of reaching the most accurate, specific diagnosis for a particular patient. ICDA is purposefully a "lumping" code system. Moreover, the ICDA categories and subcatetories are numerically coded, using a strict hierarchical structure. Hierarchical code schemes are inherently inflexible. Therefore, regrettably, ICDA cannot be developed into a medical lexicon.

Forty-five years ago, when it was planned, the Standard Nomenclature of Diseases and Operations [6], was an imaginative vocabulary with a two-faceted terminology-code scheme. After two decades, the basic logic, as well as the hierarchical code scheme, grew obsolete. The system lost both its internal consistency and its two-component paradigmatic logic. SNDO is irreversibly outdated.

The Systematized Nomenclature of Pathology (SNOP) [7] is a four-faceted vocabulary based on a linguistic analysis of diagnostic terms. SNOP was compiled with the unit record system in mind, exclusively for the pathologist. SNOP is a stock of linguistically simple, root terms which in proper paradigmatic combinations can represent compound terms (e.g., appendix + inflammation = appendicitis, colon + inflammation = colitis). The paradigmatic capability of SNOP rests more on linguistic rules, than on medical semantic proximities. Therefore, SNOP is fundamentally in conflict with the concept of a deterministic, electronic information system. Moreover, SNOP is coded in a highly hierarchical code scheme which is rigid, and therefore, unsuitable for growth. A scientific vocabulary with such a tight code scheme has a brief useful life. New knowledge may emerge unpredictably, wherever research has been productive. In such new areas, clusters of new terms are generated, whereas other terms lose meaning. Even worse, the meaning of certain terms may change. The latter entails extensive re-coding, particularly if the term with the changing meaning has a higher level position in the hierarchy [8].

It is regrettable that none of the currently available, formally published medical vocabularies can be used for computer-oriented transformation of natural language medical records. They are neither adequate to capture the semantic precision of the clinical language, nor compatible with the criteria of modern information processing technology.

4. Developing medical data systems

Today, there are more than 3,000 data centers in the U.S. handling medical data. Recent increases in governmental involvement in financing medical care will further encourage the use of computer technology. There is, however, a growing gap between the burgeoning data centers and clinical medicine. Most of these data systems were created outside the mainstream of clinical medicine. Many of these are insurance claim handlers, automated billing systems for hospitals, or statistical work for the health departments. It is regrettable that the health industry applies such vast computing power to process inaccurate medical data. This unhappy situation is further worsened by grossly inadequate coding schemes. Better planning and the involvement of clinical medicine could improve the quality of the data and the accuracy of coding.

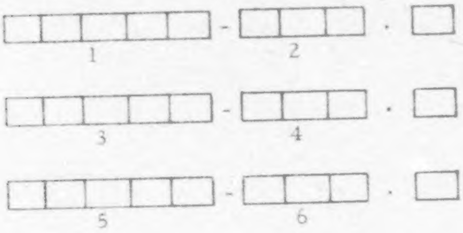
5. Confidentiality of medical information

Clinical medicine has tolerated the increasing automation of clinical data, but with a growing concern about the risk created by automation. The Medical Society of the State of New York was the first medical organization to act by recently endorsing a set of "Ethical Guidelines for Data Centers Handling Identified Medical Data [9]. If these Ethical Guidelines are honored, medical information systems can now be developed with the trust and the endorsement of clinical medicine, since in these Guidelines the humanitarian values of medicine are translated into the language of information sciences and technology.

6. Medical Data Element Dictionary

In accordance with the Ethical Guidelines [9], data systems handling medical data "shall provide assured confidentiality, and they shall conserve the integrity" of the medical records. The key to this task is an extensively documented, comprehensive Medical Data Element Dictionary (MDED). The purpose of a data center's MDED is to explicate the pertinent aspects of each medical datum accepted by the data center for storage, processing, and release. MDED is essentially a catalog of the various types of data in the information system, along with a comprehensive descriptive characterization of each datum. A data security oriented MDED should formally document each datum as follows:

Heading	Example/Explanation
NAME OF DATUM	Admitting diagnosis
OTHER NAMES	Tentative diagnosis on admission Working diagnosis upon admission Dx of admission
DATUM GENERATOR	Attending physician only!
DATA CODE	(Note: In our operation, we use two code systems: One is the specific code, a five-digit pure alphabetic code value, the other is the category code (ICDA-8)
SOURCE OF CODE	The specific codes were published by the Journal of Clinical Computing [10]; for category coding we use ICDA-8 published by U.S. Public Health Service
METHOD OF CODING	Automated letter-by-letter match encoding-decoding
DEFINITION OF DATUM	The medical term is either a diagnosis or a problem which identifies the reason for hospitalization

REMARKS	The Admitting Diagnosis is usually a single term, but occasionally, the physician may record several terms. The field can accomodate up to three admitting diagnoses. If (very rarely) the source document (Patient Registration Form) would contain more than three terms, only the first three are used
FIELD SIZE	 1-3-5 are five-position alpha codes 2-4-6 are ICDA codes
FIELD CHARACTERISTICS	(1) Some valid ICDA codes have only three code positions; (2) Some ICDA codes contain alpha characters; (3) If patient has no diagnosis (e.g., dead on arrival) specific code should be 00000, and the space for ICDA code should be 000.0; (4) Record is rejected if this field is vacant; (5) Record is rejected if either the specific code or the group code is invalid (cf. Medical Lexicon and ICDA code tables)
UNITS	Not applicable in this example (Note: units are defined here, such as mm Hg, lb, mg, cm of water)
ACCURACY	This is the most accurate datum at the time of admission (Note: generally, this segment should define the a priori expectation concerning the degree of medical-legal-scientific evidence value of the datum)
RELIABILITY	In the source document the physician's signature must follow the datum

<table>
<tr><td>USE OF DATUM</td><td>
<table>
<tr><th>Program No.</th><th>Format</th><th>Name</th></tr>
<tr><td>(1) 63</td><td>natural language</td><td>hospital census</td></tr>
<tr><td>(2) 73</td><td>coded</td><td>notification of Medicare</td></tr>
<tr><td>(3) 82</td><td>coded</td><td>daily admissions</td></tr>
</table>
(Note: This segment must be kept current. The list must be comprehensive, enumerating all programs which include the datum, as well as the frequency of usage. The program library provides further information as to authorized receivers of reports, authorized staff members to modify the program, and the data security officer's last checking of the programs involving the datum)</td></tr>
<tr><td>OWNER</td><td>Patient is the data person (Note: this segment should also refer to the password directory controlled by the authorization system and by the data security officer)</td></tr>
<tr><td>AUTOMATIC SAVE?</td><td>Yes
(Note: if the answer is "yes", (a) the datum is copied on a backup medium, e.g., microform; (b) the datum is protected from any subsequent correction or modification. The exact location of the data field on the Source Document and the location of storage must also be stated in this segment</td></tr>
<tr><td>SENSITIVITY</td><td>"Second class"
(Note: The class assigned indicates the degree of difficulty required to access the datum without proper authorization. Periodic checking of the actual difficulty must be recorded here.)</td></tr>
</table>

7. Need for coordination

The interface between clinical medicine and information technology requires an effective Medical Lexicon and the development of a purposeful MDED. Both may evolve spontaneously if adequate time is given. The acute problem is the prevailing information crisis. The risk is that in the name of expediency and economy, medical data systems may evolve without the necessary moral obligations as specified in the Ethical Guidelines. This may be simpler, and perhaps more expedient, but not prudent. Cherished social values may be permanently lost. The man-machine interface should receive high priority. A highly visible multi-disciplinary team should be created, charged with further developing the national Medical Lexicon, and continuously upgrading it. Another multidisciplinary team should encourage the development and maintenance of MDED by all data systems handling medical records. State and federal agencies should regulate, accredit, and regularly inspect medical data systems. The humanitarian values should be conserved. There is no basic conflict between human desiderata and medical information systems, unless the technology becomes autonomous, dehumanized. It is the challenge of this decade to obtain the benefits and to minimize the risks.

8. References

[1] Gabrieli, E. R.: Documentation of Laboratory Data; in Clinically Oriented Documentation of Laboratory Data, E. R. Gabrieli, ed., Academic Press, New York, 1972,

[2] Gabrieli, E. R.: Potential of Medical Computing, J. Clinical Comp., Buffalo N. Y., 1975;

[3] Gabrieli, E. R.: A National Obstetric Information System - A Proposal; J. Clin. Comp. 4:359-366, 1975;

[4] Eighth Revision International Classification of Diseases Adapted for Use in the United States, U.S. Dept. of Health, Education, and Welfare, U.S. Government Printing Office Publication No. 1693, 1968;

[5] Hospital Adaptation of ICDA (H-ICDA-2) Commission on Professional and Hospital Activities, Ann Arbor, Michigan, 1973;

[6] Standard Nomenclature of Diseases and Operations, fifth edition, ed. E. T. Thompson and A. C. Hayden, McGraw-Hill, New York, 1961;

[7] Systematized Nomenclature of Pathology College of American Pathologists, Chicago, Illinois, 1965;

[8] Gabrieli, E. R.: Coding of Medical Data; J. Clinical Computing, Buffalo, New York, 1975;

[9] Gabrieli, E. R.: Ethical Aspects of Health Records, J. Clin. Computing 4:283, 1975

[10] Gabrieli, E. R.: Design of a Medical Lexicon; J. Clin. Computing, Buffalo, New York, 1975.

Principles and Concepts
# of Data Resource Management
System Development

Aaron Hochman

DoD Logistics Data Element
Standardization & Management Office (LOGDESMO)
Office of the Assistant Secretary of Defense
(Installations and Logistics)
Alexandria, Virginia 22332

Identification of basic concepts for data resource management systems including visualization of the objectives, operating environment, subject matter content, functional activities and organizational relationships of such systems; and

Identification of the benefits expected to accrue from development and implementation of these systems.

Key words: Applications; change control; data element standardization; data fields; data relationship; data representation, data resource management; documents; files; formats; forms; management entities; management principles; plans; processes; programs; records; reports; system concepts; systems.

1. Introduction

This paper is generally intended to provide an overview of the development of systems dedicated to the management of data and information as resources in a manner similar to that employed in the management of other resources such as manpower, materiel, facilities, money, etc.

Specifically, this paper addresses:

A. An outline of recognized problems and needs for their resolution within the data and information systems community of the Federal Government;

B. An enumeration and discussion of fundamental management principles to be considered in developing data resource management systems;

C. Identification of basic concepts for data resource management systems including the objectives, operating environment, subject matter content, functional activities and organizational relationships of such systems; and

D. Identification of the benefits expected to accrue from development and implementation of these systems.

2. Background

2.1. Data System Development Trends

Characteristically, the growth of data systems has taken place primarily along functional lines, e.g. inventory management; payroll; financial management; etc. These were the areas in which the problems were most pressing and immediate and the need for more current and complete information was recognized and pursued vigorously.

The goal in each of these areas has been the integration of information. However, success has been limited to the integration of information in individual functional systems. Integration and interrelationship of the information of several systems has been slow in coming to fruition. This is understandable since interest was focused on the operations of individual functional areas and information was oriented to provide support to these operations. Relationships between functional areas and the consequent relationships between associated information were considered only incidentally.

Until recently, middle and top management had little appreciation of the potential for developing intercommunication links between information systems by using common data terms or words as a bridge. However, the rapidly evolving trend toward development of "uniform automatic data processing systems" has served to highlight a critical requirement for the integration of information within such systems.

Additionally, as more and more systems were developed and automated, it became apparent that automatic data processing equipment (ADPE) offered its users a great opportunity to make multiple use of a single document, report, word or information. Information deposited in a data bank can be withdrawn from any operational or functional perspective, thereby precluding the need to collect the original information more than once. In this process, ADPE has provided a greater capacity for the correlation of data from one system with data from another to achieve a synthesis of information that will assist management in making more timely and accurate decisions.

2.2. Data Element Standardization

With awareness of the data correlation potential of ADPE, also has come the realization that, to achieve the full potential, data terms in the information system must be uniformly identified and defined so that common elements can be isolated and standardized.

Since data are developed primarily to satisfy the requirements of individual functional systems, the same units of information are quite often found to be coded and used differently in each system. Even when such items are identified by grouping them as data items under a data element, ambiguities are still encountered when different systems employ the same data element with different combinations of data items under it. Instances are also found where the same data items are coded differently in each system. Additionally, items of information which have completly different meanings are coded identically.

A number of management programs have been established within the U. S. Federal Government to address the standardization of data elements. For further details on these programs, one should refer to the Report to the Congress by the Comptroller General of the United States dated May 16, 1974 entitled "Emphasis Needed on Government's Efforts to Standardize Data Elements and Codes for Computer Systems" (B115369) and to Federal Information Processing Standards (FIPS) Publication 28, dated December 5, 1973 entitled "Standardization of Data Elements and Representations."

2.3. Evolution of Data Base Management

As the standardization of data elements progressed, it became evident that the data element, as a fundamental unit of information, could be employed to connect the various facets of a data system thus providing a basis for data base management.

The rapid development of highly sophisticated data base management technology and applications thereof reflects a recognized critical requirement for effecting management control over the identity of data elements and the tracking of these elements to their location within the physical and logical structures of the system as well as displaying critical relationships between the elements and their use.

Data base management system technology is generally limited in use to automated data processing systems. However, it should be noted that the inherent logic of data base management could just as well be applied to a data base configuration of any type, not limited to ADPE. In this regard, a data base management system can be considered to be an ADPE System Data Resource Management System.

2.4. The Data Resource Management System (DRMS)

For the purpose of this paper, a data resource management system is defined as "a combination of management resources (e.g. man; machine; materiel; facilities; data/information; etc.) organized in a specific configuration for attaining the objective of managing data as a resource as opposed to other types of systems which address the management of other resources."

A data resource management system (DRMS) may be manual or automated or a combination of both. However, the volume of effort required to maintain such system generally dictates the use of ADP equipment. The DRMS functional scope generally encompasses, but is not limited to the establishment and maintenance of controls over requirements determination, identification, cataloging, standardization, acquisition, storage, distribution, display and disposal of information relating to data employed within one or more management systems.

A management system represents the sum of management resources organized in a finite configuration to achieve fundamental management program objectives. Its functional description includes all of the life cycle events required in the performance of program missions regardless of whether they are performed by man or machine.

The data resource management systems discussed herein include the exercise of controls over data employed not only in events accomplished by the ADPE but also in those events accomplished by human resources.

3. The Need for Data Resource Management

Continuing Congressional and General Accounting Office (GAO) criticisms of the Executive Branch of the U. S. Government have stressed a continuing need for compatibility, interface and integration in the development and operation of management and data systems. These criticisms generally identify a significant commonality of functions and a marked similarity of systems within the Federal Government Community. There is a critical need to identify and correlate and eliminate/prevent unplanned redundancy and inconsistencies in these systems.

The tremendous expenditures required to develop and maintain management and data systems in the current financially austere environment dictate that a concerted effort be made to attain a greater efficiency at a lesser cost.

The costs associated with the acquisition, handling, processing, and distribution of data represent a significant portion of these expenditures. In this regard, it is projected that as little as a one (1) percent improvement in the handling of data and information within the Federal Government would accrue savings of hundreds of millions of dollars.

The capability for effectively operating a management or data system is directly related to the effectiveness of the management of data relating to the system.

The capability for attaining maximum utilization of available data and information resources is directly related to the effectiveness of accession to such data and information. Such capability is currently impeded by the fragmentation of data controls and non-utilization of available data resource management technology.

The capability for determining the impacts of anticipated, planned or approved changes in systems and for implementing such changes is likewise hampered by the lack of a data resource management mechanism. In this regard, it is envisioned that many projected changes would be cancelled or altered if the full impacts of such changes were known initially.

The recently issued Public Laws related to "Privacy of Information" and "Freedom of Information" have further served to focus attention on statutory requirements for the management of data as a resource.

For these reasons, among others discussed in the remainder of this paper, it is considered imperative that the top and middle management echelons of the Agencies of the Federal Government seriously consider application of the data resource management system technology described herein.

4. Fundamental Management Principles

In developing a data resource management system it is essential that the following management principles be considered:

A. "An improvement in one facet of management does not necessarily improve total management"

1. The key factor in this principle is recognition of the end objectives of the organization as opposed to the end objectives of subordinate management programs and organizational sub-elements.

2. A data resource management system must be keyed to the primary objectives of the organization rather than merely responding to a compendium of subordinate objectives.

B. "Effectiveness of functional mission performance is directly related to effectiveness of data and information management"

1. Data resource management provides critically needed access to required information and reduces the time and effort required to find such information.

2. Such reduction in time and effort contributes to increased efficiency of the system users in performing their mission assignments.

3. Collaterally the availability of required information permits its reutilization and thus avoids the necessity for duplicative data procurement and/or development.

C. "Effectiveness of data resource management involves efficiency in providing the data customer with the information that he requires, on time, in a form that he can understand without requiring unnecessary effort on the part of the customer"

1. The purpose of the data resource management system is to provide necessary information support services to system users.

2. The system customers' acceptance and use of the system is contingent upon the responsiveness of the system to users' needs.

3. Such responsiveness is directly related to the facility of accession, timeliness of retrieval, and usefulness of the information provided.

4. Futher, it is well recognized that the user of an information retrieval system will tend not to use the system if retrieval involves any significant level of work to be performed by the user.

D. "Data is a critical resource and deserves the same kind of management disciplines as those employed in the mangement of other resources"

1. Data and information are fundamental elements of any system whether it is a management system, a weapon system or a data system.

2. Data and information are subject to the same life cycle events as other resources (e.g., requirements determination; development; identification, cataloging; acquisition; storage; maintenance; issue; utilization; disposal; etc.).

3. Data and information are not normally expended when used and in this sense, they provide a potential for reutilization not common to other resources. It is this reutilization potential that dictates the application of data resource management particularly in an environment where the costs of doing business are overwhelmingly dominated by the costs of human resources who in turn are the "consumers" of data and information.

E. "Data can be managed on either a commodity or system oriented basis or a combination of both"

1. Commodity oriented management is essentially concerned with what the data are, independent of application.

2. System oriented data management is primarily concerned with how the data are applied.

3. Data resource management provides the critically required correlation of the substantively different forms of data management.

F. "Computers are intended to be labor saving devices"

1. The amount of effort expended by the human in relation to the computer must be kept to a minimum.

2. The principal factor is: "Say not what you can do for the computer; say what the computer can do for you."

5. Concepts

5.1. Objectives of the DRMS

Generally speaking the DRMS is intended to provide the most effective and economical information support of personnel engaged in planning, developing, designing, implementing, maintaining, operating, utilizing and managing automated, mechanized and/or manual management or data systems.

Specifically, the DRMS is designed to:

A. Improve the management of resources at every level through reduction of unnecessary redundancy and elimination of unplanned inconsistency.

B. Reduce costs of data systems design, development, programming and operation.

C. Assure effective management of existing and planned data bases.

D. Facilitate interface and integration of data and management systems and the interchange of data between and among such systems.

E. Provide a common base of standard data elements and related features for use throughout the community of users served by the system.

F. Permit more efficient determination of the impact of anticipated, proposed and/or approved changes by those organizational elements planning, administering and maintaining programs and systems employing the data base content which the DRMS addresses.

G. Establish appropriate monitor controls and surveillance as required by public laws relating to "Privacy in Information" and "Freedom of Information."

H. Provide the participating system users with required information support services emphasizing giving the customer what he requires, on time, in understandable form and format and without requiring any significant level of work to be performed by the user.

5.2. Operating Environment

The DRMS is designed to operate in an environment comprised of a hierarchy of management entities within one or more specified organizations.

An organization is definable in terms of the functions it performs within given subject matter areas.

A major management entity is the documentation used to identify, characterize and prescribe the use of other management entities (e.g. Directives, Instructions, Regulations, Manuals, Handbooks, etc.).

The other principal management entities represent the structured configurations of management plans, management programs and systems.

For the purposes of this paper, a system is considered to be a combination of management resources (human resources, materiel resources, real property resources, financial resources, data and information resources and natural resources) organized in prescribed proportions and configured to attain identified goals and objectives. In this context, the term system is not limited to automated systems.

Systems, in turn subdivide into applications (subsystems).

Manual applications are further sub-structured into manual operations and manual procedures whereas automated systems are identified in terms of processes and automated processing programs.

Regardless of whether a system is automated, manual or hybrid, the physical management entities relating to the system and its subdivision can be uniformly identified as:

A. Forms - a selected medium of data recording with blocks and/or columns of prescribed data requirements but without actual data entries (blank form).

B. Format - the tutorial prescription of data requirements to be entered on a record, document or report but without actual data entries or provision of a medium for data recording.

C. Files - a collection of records.

D. Records - a unit component of physical file data storage presenting a configured array of data entries in accordance with a prescribed format recorded on a prescribed form. Manual records generally include both the data requirements and the data entries applicable to such requirements. Machine sensible records most often convey only the data entries with the data requirements implied through position location of the data entries within the record.

E. Documents (Inputs) - a structured string of data in prescribed form and format used as an input from human or machine source designed to trigger specific actions within an addressee system or subsystem.

F. Reports (Outputs) - the structured product display in prescribed form and format intended to communicate required information to a system user.

These physical entities include data fields (blocks) which represent the data requirement with data entries in the case of records documents and reports and without data entries in the case of forms and formats. At this level the data field (block) constitutes a basic data unit within a specified system.

Data Fields (Blocks) are further identifiable in terms of applicable data elements which constitute the basic data unit between and among various systems.

The data field (block) identification (identifying number and/or name) may be likened to the reference number and name assigned to a particular piece or part of an equipment whereas the data element using the same analogy, may be likened to the uniform item identification and stock number assigned to an item of supply based upon what it is rather than how it is used.

It should be recognized that within a specific system, the data field (block) quite often is identified as a "data element." However, within the concept of the DRMS, the data element represents a more universal unit of data and is identified apart from the data field (block).

The data element, in turn, relates logically to its component data items and their representations in literal, abbreviated or coded form.

5.3. Subject Matter Content

The data base of the DRMS encompasses data element identification and characteristics information as well as correlative information concerning data element applications.

The DRMS is designed for commodity oriented management of data with system oriented applications cross-referenced as attributes.

The data element serves as the basic unit of information in the DRMS. Attributes of the data element are introduced into the system records from the various authoritative issuances, functional procedures, data system physical and logical structures, reports and forms contingent upon the scope of the system. Additive control attributes are also added as tags or labels as required for security, privacy or freedom of information considerations.

It is emphasized again that the data content is not limited to data system structures but also includes all logical structures of the management system external to ADP systems.

5.4. Functional Activities

The basic functions of the DRMS are those of any conventional data bank. (e.g., data acquisition; input; storage; retrieval and output of products display).

A major feature of the DRMS is the process of associating related sets of information so that they may be displayed when required.

The DRMS, as projected, will provide two basic modes of output identified as push and pull.

A. The push mode provides for cyclic output of formatted product reports, listings or publications with associated cross-reference indices. When the size of these outputs warrant it, the products can be produced directly on microfiche cards thus enabling more timely and much more economical distribution to the customer. In this regard, it should be noted that the COM equipment employed to create the microfiche card directly from magnetic tape, operates at least forty (40) times faster than the fastest line printer. Each microfiche card contains 280 page images, 279 of which are text and the 280th page image which is used to index each image subject content (first line entry) to a coordinate (row and column) location. Additionally, the major heading of the card in oversize lettering is produced during the same process. The resulting products are extremely low cost, not only in production, but in handling and storage space as well. The conventional trade-offs of producing or not producing a product because of its size must be reconsidered in the light of this new technology. The costs of copying microfiche cards is also exceptionally low. The need to use a change bulletin or supplement to avoid long tedious listings must also be reconsidered. Finally, the reduction of lead time to produce a publication including elimination of all intermediate handling normally required when hard copy is involved has increased the currency of the information being distributed to a point where most customers needs are satisfied except for a select few who must operate on the basis of immediately current data.

B. For this selective range of users, the pull mode provides for on-line interactive retrieval. The user is not constrained by cut-off dates for publication or by the limitations of list type indices. The interactive retrieval features enable browze search employing descriptor sets on an and; or; or and/or basis. The user is provided with visibility of available sets of related information current as of the time of interrogation. The response times for retrieval are normally measurable in fractions of a minute and in some more sophisticated systems in terms of seconds. It is important to reemphasize that this mode should be restricted to only those users who have a demonstrated need for absolutely current information.

5.5 Organizational Relationships

The organization designated as responsible for managing the DRMS is a data commodity manager.

The inputters, other than the responsible organization are generally systems oriented, i.e. concerned with how the data are used.

It is critical that the interface of the inputters, with the DRMS, be effected with minimum disruption to ongoing system oriented management.

The users must be identified, their requirements identified and their interface with the DRMS established on either the push or pull mode. To assure user acceptance and use of the DRMS output products, it is essential that such products or displays be organized to satisfy user requirements and that information accession be simplified to the extent possible.

6. Anticipated Benefits

Installation and implementation of the DRMS will provide major benefits including:

A. Increased efficiency on the part of functional program management and operating personnel in the performance of their assigned missions.

B. Reduction of costs of developing, designing, operating and maintaining data and management systems.

C. Comprehensive visibility of the relationships between programs, systems and their data thus providing an invaluable management tool for system development planning.

D. Increased capability for the reutilization of available data and information resources.

E. Provision of an effective mechanism for planning and controlling the introduction of changes into data and management systems by means of impact analysis.

Above all, the principal benefit of the DRMS is that it offers an effective means of "providing the customer with what he needs, on time and in an understandable manner without imposing any significant workload on the customer." The DRMS approach, as observed to date, is considerably more effective and economical than other techniques in use throughout the Federal Government and its use is commended to the management of all Federal Agencies.

The Design of Data Elements:
A Data Base Perspective

M. A. Huffenberger[1]

Chemical Abstracts Service
The Ohio State University
Columbus, Ohio 43210

Data elements are a major concern at Chemical Abstracts Service (CAS), where the business is "information." The role of CAS as a major information processor has dictated the development of long-term objectives for information storage and retrieval. Based on these objectives, the principles by which data elements should be designed are described. Current CAS practices in data element and file design which contribute to standardization and the interchange of computer-readable data are discussed.

Key Words: Data base; data base concept; data base manager (DBM); data base management system (DBMS); Data Dictionary/Directory (DD/D); data element; data independence; data integrity; documentation; programming viability; Standard Distribution Format (SDF); Standard File Format (SFF); statistics.

1. Introduction

Data elements are the *sine qua non* of an information processing system -- the sum of its input, its substance, and the source of its output. For this reason, data element definition guidelines significantly affect the information processing environment itself. This influence may extend over a number of years, as in general it is the means by which data is processed rather than the nature of the data that changes with time. For example, while the representation of a chemical substance by a structural formula is over a century old, three generations of the CAS Chemical Registry System have been developed in the past decade to process fundamentally the same information (including structural formulas). Thus, the role of defining the data elements within a data base is a profound undertaking.

The modern expression of the data element concept began for CAS in 1966. At that time, a program was chartered to collect, identify, and organize data items to be processed by the CAS information-handling system. CAS also devised a new scheme of file organization, Standard File Format (SFF), which was designed to provide some measure of data independence in data element storage and to allow flexibility in the structure of file segments by accommodating both fixed and variable length data elements, which occur one or more times, or are indeed absent altogether when they are optional. [1][2] The necessity of this flexibility follows from the extreme variability of the CAS file environment. Today, CAS files involve millions of occurrences of the 2000 different defined data elements, including abstract and index text, chemical structures, article and journal titles, chemical substance names, chemical properties, and subscriber, personnel/payroll, and accounting information.

---

[1] System Engineer

[2] Figures in brackets indicate the literature references at the end of this paper.

By establishing consistency in data element definition, CAS was also able to standardize storage and retrieval software. Thus, powerful processing, storage, and retrieval tools have been developed to complement the large data bases which have developed over the years. [2] To accommodate the increasing growth and complexity of its computing environment, CAS has established a unit within its Research and Development Division, the Data Management Services unit, to handle a continuing program of data element and data base management. This unit also participates in the development of data base management concepts and tools.

2. Significance of Data Elements

Before describing data element design at CAS, one must define the term "data element." For the purposes of this discussion, a data element will be defined as the elementary form in which data is stored, retrieved, and processed by data management software. No smaller unit of data is uniquely identified and formally documented as an independent entity in the CAS data processing system.

Such a definition allows considerable latitude in establishing data element structure. A data element may be a single item such as a country name, a simple chain of items such as a date shown as YYMMDD, or an array of items. An example of an array would be Mailing Address, including Name, Street and Number, City, State, and Zip Code. Such data elements consist of diverse collections of items, usually in a fixed format design.[3] These complex forms may be thought of as "pseudo-segments" because a multiplicity of data items is retrieved in a single data-locating operation. Compound data elements, however, may tend to defeat the purposes for which data elements are defined as will be discussed in section 4.

The primary value of defining data elements is to provide a unique identifier for a given data entity. These global identifiers allow the various components of the information processing system to correspond and interact with one another through standard interfaces. A lack of unique identification necessitates interface programs to translate identifiers for the various program components of the processing system. This lack also leads to redundancy and spurious identification of data fields as new application systems redefine and store a given item of data under various labels.

The complement to unique labeling is a standard representation for a given data entity. For instance, a date can be displayed as YYMMDD, DDMMYY, Julian, or otherwise, but it should be stored in a data base in only one of these possibilities. If an application program requires an alternative format, it should explicitly state so in the retrieval process. A single generalized translation routine can then be invoked. The alternative to a standard representation again is an undesirable battery of interface programs.

The unique identification of data elements and standardization of their formats offers several advantages. Standard documentation, text and title indexes, and cross-references are possible. Different categories of data elements can be recognized, and guidelines can be prepared to encourage the most desirable styles of definition. Consistent processing of similar types of data can be initiated with generalized editing routines. For example, all data elements specifying money amounts can be edited as a class having numerical values with two decimal places.

All these advantages support the data base concept by facilitating data interchange among components of the information processing system and promoting the design of shared files. Progress toward the concurrent goal of data independence is also furthered by the formal definition of data elements since the encoding of parochial and nonstandard data element formats into programs is discouraged. These topics will be discussed in the following sections.

---

[3] In some terminologies this type of data element would be referred to as a "group" element; here they will be called "compound" data elements.

3. General Principles

Before determining the principles and practices for data element definition, one must establish well-defined objectives to be met by these definitions and determine how these objectives may be achieved.

Thus, the definition process must be examined in terms such as the data base concept, data independence, data integrity, adequate documentation, and programming viability. These goals provide a framework within which to develop guidelines and procedures for data element definition.

3.1. The Data Base Concept

The data base concept asserts that there exists, for each enterprise, an accumulation of data which is pivotal to its operation. The data may be personnel/payroll data, shipping and billing records, inventories, or accounts. Such a body of data is called "the corporate data base." In a practical sense, this does not imply a monolith; several application data bases, all defined according to the same principles and using a unified data element set, may exist. For example, CAS has a Chemical Registry Data Base and a Publication Data Base, among others.

The data base concept implies that the description and treatment of data should cease to be oriented toward specific processes and that the intrinsic value of the data and its universal character exceed the scope of a single application system. Furthermore, a data base is not specifically owned or controlled by any single user, but may be updated by or supply information to several users independently. [3,4,5]

These principles require that data element descriptions be closely reviewed by a central authority to prevent overly specific titles or descriptions, which could exclude possible new users of the data base or require expensive changes to achieve generality. The data element definition should depict an entity of data, not a processing context. To this end, generalized data elements can be an invaluable aid. The context in which one of these elements appears frequently is enough to eliminate possible ambiguity. For example, rather than defining Shipping Zip Code and Billing Zip Code, it is sufficient to define Zip Code and use it in shipping and billing records which specify what type of zip code is involved.

For users submitting drafts of new data elements, the possibility that similar definitions already exist must be investigated. To make this search and review practical, each data element must be clearly titled and classified according to the type of data contained. Once a data element has been established, new elements which represent identical material should not be defined. Sometimes in fact, not only are new data elements not required, but the desired data is already on the data base! To avoid establishing separate data elements with similar definitions requires a knowledge of the constituent data elements for each existing file. Such a capability will also indicate whether data occur (or will occur) redundantly on the files. For example, processing tractability may require that a given value of a data element be present on several files in the data base.

Storage modes and formats must be carefully determined. Certain storage modes may be totally inappropriate for a specific data element and its likely use. For example, on the IBM 370 computer, statistical values stored in EBCDIC must be converted to Packed Decimal or Binary before arithmetic is performed. A representation or format of a data entity must also be acceptable to a wide audience as standard representation is a primary goal of data element definition.

Data should be stored only in "whole" pieces although less might suffice for a specific application. For example, consider a system in which personnel data are input. Both Social Security Number and Name are entered, but only eight characters of the name are retained. While this abbreviated form saves storage and suffices for sorting into recognizable name order, a future system to print payroll checks will be unable to use the abbreviated name and would redundantly input this information itself.

In summary, the data base concept argues for data element definitions amenable to visibility of the data entities, standardization, and generality for flexible usage.

3.2. Data Independence

Complete data independence means that applications utilizing the data base depend only on its contents -- and not on the data organization, formats, storage modes, or storage locations. [6] Several implications follow from this definition. With data independence, data elements may be used by various programs in a multitude of formats and storage modes, regardless of the stored form. Application programs may receive "derived" data as if this data were stored already in the data base. Data may be presented in different sort order than that in which it is stored. The data base may be restructured without programming changes, and programs may be added or changed without modifying the data base. Although these capabilities have been grouped under the term "data independence," they are clearly implied by the data base concept, which asserts that data no longer belongs to individual programs or systems and thus cannot be too specialized or interwoven in the logic of application programs.

To clarify the nature of data independence, we need to examine the parameters describing a data element on the data base (table 1).

Table 1. Parameters Characterizing a Data Element on the Data Base

Parameter	Explanation or Example
Location	Address (file, segment, location within segment)
Encoding Scheme/Storage Mode	Binary, Packed Decimal, ASCII, EBCDIC
Character Set	numerics only, {A-Z,/,.}
ID Number	0859
Range of Values	0-9999, 1966-1976
Physical Size	1-15 bytes
Representation/Format	J. L. Smith vs. Smith, J. L. MMDDYY vs. YYDDD 2261 vs. 2,261

The number of parameters necessary to characterize a data element may be installation dependent -- fewer or more parameters may be necessary in different environments. These parameters will be predefined (implicit, fixed) or carried dynamically (explicitly stored with the data or in specially designated control areas). The fewer parameters that must be compiled into an application program in order to process a data element, the greater the program's degree of data independence. One need not presume that freedom from each of the parameters is equally practical or desirable; location independence may be better than range independence. At CAS storage mode independence seems to be more practical than representation/format independence; thus, this paper advocates standardizing representations, yet describes flexibility in the use of storage modes.

What enables an information processing system to achieve data independence? Inherent in data independence is the ability for the form of stored data (sometimes called "physical data") to differ from that of the data presented to application programs ("logical data"). In general, the application programs should specify the desired data elements, data element formats, and storage modes, so that these data elements may be assembled from the data base as required. Data elements submitted to be stored must be channeled and converted to satisfy the format, storage mode, and content requirements of the data base. Obviously, a mechanism is needed to perform these tasks -- a collection of programs called a data base manager (DBM) is utilized. A repository of data base contents, relationships, and specifications is also required. This role is fulfilled by the Data Dictionary/Directory (DD/D), a collection of authority files which characterize a company's data bases. [6,7] The "characterization" of a data base includes, for data elements, textual descriptions, cross-references, statistics, physical specifications, access restrictions, and where-used information. The DD/D becomes part of the corporate data base, and its characterizations help make the informational data bases self-defining.

The access software (DBM) of the data base management system (DBMS) at CAS is called FIDO. [8] Several features of this software support data independence, including retrieval by index files (a technique also known as "partial inversion"), segment validation (which certifies that all required data elements are present), ICL/XCL processing (inclusive/exclusive lists of data elements to be operated on at one time in a segment), and automatic storage mode conversion.

Establishing data element definition practices in accordance with data base concept guidelines will further support data independence. Using standard representations and general, data-oriented (rather than process-oriented) descriptions will discourage high dependency between specific programs and the data base. For example, CAS does not consider storage mode differences among otherwise identical data elements to require separate data element ID numbers. By programs taking advantage of storage mode specification stored along with the data element itself and available conversion routines, redundant definitions can be prevented. Some care must be exercised to assure that two versions of a data item differ only by a standard storage mode conversion when the above rule is applied.

Example 1. The number "123" may be alternatively encoded as

F1 F2 F3 = 7B[4]

EBCDIC Binary

but a date written YYMMDD cannot be interpreted as MMDDYY regardless of storage mode (i.e., these are two separate representations or formats).

Dependency can also be discouraged by limiting the complexity of data element formats. Data elements with substructure (i.e., composed of several identifiable data items) are not self-defining. That role will have to be assumed by application program code unless there is a DD/D to supply the DBM with a structure algorithm, and the DBM in turn derives individual item data elements. More specific treatment of this problem will be discussed in section 4, "Practical Aspects -- Storage and Processing Requirements."

Data independence is an abstract goal, but the advantages are real. The expenses of program and data base modifications can be significantly decreased by judicious use of its concepts.

3.3 Data Integrity

The tenets of data integrity are that the data be consistent, both with associated data and with documentation, and that the data be correct within defined limits of precision (i.e., the values lie within a valid range and use the proper character set). The consideration of consistency and correctness within the data base should begin at data element definition time.

Creating a data element which is unnecessarily difficult to input, process, and validate can endanger integrity. For example, State Name can be input expeditiously as the formally accepted two character abbreviation; misspelling MS is not so likely as misspelling Mississippi. Data elements with intricate internal structure may burden application programs with complex logic and make maintenance activities error prone.

---

[4] These values are written in hexadecimal representation for convenience.

In the State Name example above, an expansion table would be necessary for display purposes, and guaranteeing the completeness and accuracy of such tables is important. Tables for character conversion and code expansion should not be imbedded within application programs throughout the processing system. Instead, a common table which is part of the data base and is independent of specific programs should be referenced. Changing these character sets or code expansions can then be accomplished as a single action. Centralized responsibility for maintaining such tables is desirable.

Some data elements can be defined as codes with self-checking properties -- utilizing a check character, such as the CODEN data element at CAS. The check character is derived from and appended to the data portion of the code. An example of a modulus 5 self-checking code follows.[5] Given a numeric code of three characters, add a check character (modulus 5) based on the sum of the original three.

Example 2.

Original Code	265
Check Character Calculation	$\frac{2+6+5}{5} = \frac{13}{5} = 3 \pmod 5$
Self-Checking Code	2653

The purpose of a check character is to avoid inadvertent miscopying of the code resulting in a valid, but incorrect value. In the above example, miscopying the 5 as an 8 yields a check character of 1; i.e., $(2+6+8)/5 = 1 \pmod 5$, which disagrees with the original resulting check character, 3. This fact indicates to a keyboarder that a mistake has occurred. Note, however, that this sample scheme is helpless against transposition errors: 265 and 625 have the same check character. The limitations of self-checking schemes, therefore, must always be examined.

Code-type elements may be defined to provide a low potential for input errors. For example, the input error rate for codes longer than five letters or six digits increases almost exponentially with increasing length. [9] In addition to length restrictions, mnemonics can improve accuracy. For this reason, mnemonics for keyboard ID's are utilized at CAS. For necessarily long codes, insertion of hyphens is suggested; social security numbers are an example of this practice.

Integrity can fall prey to invalid algorithms and inflexible code schemes which may not produce unique codes or may be too limited in scope to handle a given number or certain types of transactions. Rigorous testing of new program code and a disciplined program change environment can help to alleviate these problems. Test data by which success or failure of a program can be measured should be maintained, preferably by an authority independent of the programming task itself.

Another insidious possibility is the use of data elements not constrained within expected limits of precision or format. For example, consider a data element called File Size to contain the number of bytes a file occupies. Imagine the outcome for a file survey program that encounters the data value "UNKNOWN" and attempts to perform arithmetic. Yet this spurious value is certainly a plausible entry from a keyboard. Input-edit programs, interactive at keyboarding time, are vital to detect this type of error.

There are a number of other integrity considerations for establishing a new data element. The responsibility for input and validation must be delegated carefully so that unauthorized components of the information processing system cannot intentionally or unintentionally read or modify the data. Feedback mechanisms should be available to guarantee consistency and correctness. Validating data should involve not only input activity but also routine audits of the data base to prevent gradual degradation of its quality. This checking may be performed by a central authority or by the individual application system, but ensuring the integrity of data begins with the primary level of the data entity hierarchy -- the data elements.

---

[5] This example is deliberately oversimplified for illustration and is not suggested as a practical self-checking code. The CODEN data element cited uses a scheme more complex than this. For very high reliability and protection against a wide variety of error types, much more complexity and redundancy can be built into self-checking codes.

3.4. Adequate Documentation

Adequate documentation of a data element ensures that all pertinent facts about the element are communicated directly and concisely for both programming and non-programming purposes. Documentation services should include cross-reference lists and prompting facilities, if necessary, to guarantee that the user becomes aware of the full scope of documentation available, and of interrelationships between the data entities pursuant to his query. The Data Dictionary Directory (DD/D) is gaining prominence as a valuable tool in this respect.

Good overall documentation must reveal what a data element represents, what it looks like, how and where it is used, and contain relevant notes about its generation, validation, and history. Since the data element description serves as the focus of this documentation, its contents will now be examined.

First, the element must be properly identified. At CAS, a four character hexadecimal identification number is used. Although unique naming schemes can be derived based on data classification, an arbitrary numbering scheme is simpler and avoids the rigidity of a classification scheme that may not anticipate all future requirements. Nonetheless, data classification can be a fruitful approach to analyzing a collection of data elements. [2]

Data elements should have concise, meaningful titles, accompanied by a keyword index to facilitate searches. Synonyms and keyboard identification numbers should be included in a cross-reference list. Titles should not be more specific than the data entity they identify. For example, Payroll Master File Employee Name is an undesirable title for a data element (Employee Name) that may appear on various business and personnel files not involved in payroll processing.

A textual description of the data element is required. As with titles, the description should be definitive but not unnecessarily specific. For example, the files or programs in which a data element is used should not be listed within the description. This information is vital in overall file and system descriptions but not in the description of a data entity. This information tends to change frequently as new systems are designed, and such changes could invalidate reams of data element documentation. Where-used information can be carried in the documentation on the appropriate aspect of the system, i.e., file, program, or total system.

The physical layout of a data element must be recorded. One must know how compound elements are formatted in order to build and read them correctly. The value of simplicity in depicting the layout of a data element cannot be overemphasized; the slightest inconsistency or ambiguity in a complex definition promotes future problems.

Validating rules or editing routines[6] must be recorded. These guarantee a consistent interpretation of the format and possible values of a data element. In defining an element, the system analyst should not limit lengths and ranges of the values to those required by his own system. As much as possible, these limits should be determined by the characteristics of the data itself. Limiting the scope of data element usage by a restrictive description contradicts the data base concept. While storage mode is included in the description, it should be clear from the discussion of storage mode translations that a given element may appear in diverse modes with different coding schemes (Binary, ASCII, EBCDIC, Packed Decimal, etc.). The original specification, then, represents a "basis" mode from which others may be derived. Length specifications may also vary based on different storage modes.

---

[6] Editing routines, when established for generalized data elements, should be application independent.

Special notes and further references about the creation or processing of a data element can also be extremely useful. For example, the existence of authority tables of values should be documented. Restrictive conditions on the use of the data element (i.e., special security or integrity measures) should be listed along with a justification for these restrictions and the using or updating authority.

In addition to the information on the data element description form, several other mandatory items of information about data elements will appear in other sources. Complete documentation of a data base will link data elements to the segments, files, and processes in which they appear. At CAS, file control blocks depicting this information were introduced with the advent of the data base management system (DBMS). The structure, contents, physical specifications, and accessing lists for each file are resident on-line for use by the DBM software; this data constitutes a DD/D. Today, there are hundreds of these control blocks which play a vital role in data base documentation and integrity.

The documented relationships (in the control blocks) between data elements, segments, and files allow the DBM software to validate segments created for the data base by ensuring that required types of segments will be present on a file and that the segments will contain all required data elements. Automatic storage mode conversion is performed on data elements based on values stored in the file control blocks. Other control blocks which link programs to files afford security and integrity precautions by limiting file access to authorized programs only.

The tracing capability afforded by linking data elements to segments, files, and programs allows the impacts of data element modifications to be gauged prior to implementation. Changing the format, length, or acceptable range of values for a data element may mean program changes in several application systems. Programs can be corrected in anticipation of data element revisions rather than as a belated response to system failure.

The file control blocks have also proved to be a valuable reference tool for systems analysts and programmers who need to be aware of file structures, contents, and file-program interrelationships.

3.5. Programming Viability

Programming viability draws heavily from the topics just discussed; nevertheless it merits individual attention. Programming viability, in the context of this discussion, is a measure of the success application programs have in creating or accessing a data element. Is the data element efficient to handle? Is the associated program logic so complex as to be error prone?

Documentation often plays a key role. Of prime importance is the comprehensive and comprehensible data element description already discussed. Programmers and analysts must not subvert the documentation by deviating from published specifications. These deviations can occur in the form of new, undocumented code values appearing on the data base. Other minor deviations from allowable characters and ranges of values are unfortunately common. If changes are necessary, the documentation must be kept up-to-date.

Programming viability also depends on the generality of already-defined data elements -- can they serve in a broader context or are they too specialized? Application programmers should take advantage of standard editing routines for elements, rather than writing new and probably slightly different versions just for their own programs. Integrity conflicts may result when data elements are not subject to standardized editing.

Programming viability is enhanced by simple, unstructured data elements which require very little application program code to interpret. Thereby, the incidence of program errors is certainly reduced. Also, incorporating data structure into program code is detrimental to data independence, as discussed earlier.

Data should be stored in machine-manipulative form rather than as notes or comments. For example, consider the data element "Number of Dependents," with the following data value: "This employee claims 2 dependents." The example is informative for display purposes but is not computationally useful. It could be useful since little is communicated beyond what the data element title and the value "2" communicate. A related problem is defining a new data element to be the display form of a data element already available on the data base. Especially in on-line situations where storage space is important, enhancement of format and textual embellishment should be performed only when display or reporting is required to prevent storing superfluous material on the data base.

Programming viability also improves with each step toward a data base access system that effects data independence. The goal of independence is to remove the detailed and complex aspects of data element form, location, and control from the concern of the application programmer, yet provide effective access tools for the data he needs to process.

3.6. Statistics for Design, Monitoring, and Reference

Before deciding how to apply the principles just discussed to data element definition, one must be able to characterize the existing contents of the data base. Thus, statistics describing data elements are mandatory for good data base management. The two fundamental parameters describing data elements are length and number of occurrences, which may be denoted by a variety of statistical parameters -- minimum, maximum, median, mean, etc. These statistics must be obtained in a timely and accurate manner because they describe the actual workings within the file environment. Tools such as auditing programs are invaluable for these purposes, and quite a reservoir of audit data is maintained at CAS for on-demand reference.

Design and developmental tasks use data element statistics in several ways. In an on-line environment, the management of direct access resources requires accurate prediction of storage requirements for files. The determination of file size may also impact the assignment of storage modes and storage formats for data elements; restrictions may dictate compact storage modes. Space restrictions may also call for combining the numerous data items into several compound data elements. The impact of this compounding on data independence will be discussed in section 4. Appropriate projected or derived figures can be used to model a new data base with reasonable accuracy and to compare alternative strategies for storage and retrieval. Of course, other parameters such as number of operations required to access specific data elements and frequency of access must also be considered.

Another use of statistics is in maintaining an audit trail of processing. Consistent numbers of data element occurrences at various stages in the information processing system can be used to verify system integrity. A sudden and inexplicable change in the quantities or physical characteristics of data elements may imply a breakdown in the system. Data may have been altered or dropped; these warning signals can indicate the need to initiate specific diagnostic analyses.

Data element statistics at CAS are a valuable reference tool for various kinds of inquiries about data on the computer files. For example, how many authors (i.e., Author Name data elements) are on the Publication Data Base? What are the minimum, average, and maximum lengths of the existing names? How many authors per journal article are typical (number of Article Title elements divided by the number of Author Name elements)? Often users are not aware that data element statistics supplied by file audits can provide answers to these kinds of questions. A key responsibility for those involved in data base management is to make statistics available to users by helping phrase their questions in data element terms.

4. Practical Aspects -- Storage and Processing Requirements

The discussion thus far has focused on guidelines for data element definition and representation which support the overall objectives of the data base concept, data independence, data integrity, adequate documentation, programming viability, and the use of data element statistics in achieving these objectives. Now, one of the techniques employed at CAS to accomplish these goals will be examined.

As previously mentioned, the Standard File Format (SFF) technique for storing data elements was developed partly to provide some measure of data independence. [1,10] The specific layouts of stored segments are inconsequential to application programs. Data elements may be present or absent, and may vary in length. There may be repeated occurrences of a data element (e.g., an abstracted paper may have several authors, whose names could be stored as repetitions of the Author Name data element). Such variability in file segments does not require program changes for each segment configuration. At the segment level, files in SFF are independent of storage medium (i.e., tape or disk). Data elements may appear in various storage modes (e.g., the Registry Number of a chemical compound is in binary in one file and in ASCII in another).

These advantages are possible because SFF segments are self-defining; ID numbers, locations, and lengths of the data elements are stored in tags within the segment itself. The storage mode of each element is also contained within the individual data element tag, enabling independence of storage mode (see fig. 1).

Different subsets of the elements within a segment may be retrieved by different programs; thus, the logical and physical forms of a segment may be quite different. These qualities (i.e., binding to the data at execution time rather than compile time) are ideal in a data base environment, and experience has shown SFF to be an adequate provider -- it is an institution in CAS file design.

Standard Distribution Format (SDF), the format in which computer-readable products are issued by CAS, is a derivative of the Standard File Format (SFF) used for internal processing operations. SDF is obtained by (1) limiting the character sets used for textual data to ASCII only, and (2) eliminating certain data elements used for internal control, reporting, and record-keeping. The genetic similarity between SFF and SDF ensures that the data independence measures afforded by one are also characteristics of the other. Thus, the interchange of SDF products between CAS and subscriber systems proceeds straightforwardly.

The manner in which SFF supports data independence has been discussed. Next, one must consider the implications of using such a technique in actually building a data base. The prime considerations are the impacts on storage and processing resources, since it is virtually axiomatic that building flexibility into a data base will increase the requirements for these resources.

The cost of SFF may be examined by considering its self-defining properties, which are the measures contributing to data independence. Standard File Format segments are self-defining because of a collection of tags which identifies and points to the data contents of the segment. The storage of this information means that SFF segments are only partially filled by data content.[7] Another factor is the alignment of data contents on specific boundaries, originally due to machine access efficiency considerations. An SFF segment can be mathematically described with the following expression:

$$\text{Segment Length} = \sum_{i=1}^{n} (E_i + T_i + A_i) + 8 \qquad (1)$$

where E_i is the length of data element i, T_i is the length of the tag, and A_i is the alignment padding incurred.[8] An indigenous segment descriptor requires 8 bytes. The number of data elements in the segment is represented by n.

Studies have shown that the quantity A_i and the 8 byte segment descriptor are not significant factors in SFF storage requirements. The ratio of tag length to data element length (T_i/E_i) is most important. This ratio is typically 0.4. Figure 2 demonstrates the relationship between data element length and SFF storage requirements.

---

[7] Results obtained in 1969 [1] indicated that data content generally occupies 60% of the file space, data element tags 30%, and boundary alignment padding 10%. In spite of this, new SFF versions of files tested were not significantly larger than non-SFF versions with "untagged" data elements. This is primarily due to the fact that the non-SFF versions were forced to pad out data fields to conform to a fixed format scheme.

[8] For data elements whose data values are four or less bytes in length, the data replaces the Location and Length fields of the tag. For this "short form" storage $A_i = 0$, and thus $E_i + T_i + A_i = 8$.

Figure 2 clearly illustrates that grouping small data elements together to form larger compound data elements will require less total file space (as long as enough of the small units are always present to form compounds). This reduction in file space is because fewer tags need to be stored. File access time also is decreased by retrieving several items as a compound data element rather than requiring a separate access for each.

But compounding disrupts the self-defining aspects of the SFF data base by "hiding" items of data which should be individually labeled. Independence is adversely affected, since the internal structure of a compound element must be recognized by an application program. A possible alternative is to incorporate a compound catalog as part of a DD/D to facilitate assembly and disassembly of compound elements. Thus, an item could be retrieved by its data element identification number regardless of whether it appeared as an individual data element or as a member of a compound element. The storage and processing trade-offs of design alternatives for this approach are to be examined. As in this discussion of SFF, the commitment to data base objectives always impacts storage and processing resources.

The final consideration in defining data elements is to establish a balance between idealism and practicality. At some point, design logic confronts the mechanical aspects of computer system operation. Production schedules and economics are very real obligations for both application system designers and data base analysts. The workability of a new application system can be compromised by overindulging in data base principles to the exclusion of efficient programming practices. In general, the best approach is to start with a design which fulfills the goals suggested here and to yield to processing and storage efficiency arguments only when necessary. Some necessary "horse-trading" results between programmers, systems analysts, and data base specialists because there is no simple formula by which to compute the most beneficial design overall.

For instance, given a data element composed of byte switches used to control editing modules in an application system, should one insist on defining and storing (via SFF) each switch as an individual element? Probably not, although the goal is to simplify data elements, prevent binding to individual programs, and encourage data base flexibility. This data element is probably process specific by definition. In addition, fractionation would increase the storage and accessing overhead to an unacceptable degree. The objectives of data element design must be framed in their proper perspective.

5. Conclusion

The design of data elements is not a simple topic. The trade-offs between storage and processing efficiency and data base ideals become quite complex -- intuition, experience, and analysis all must be brought to bear when examining proposed new data elements.

Hopefully, this discussion has illuminated some of the considerations that must be made. Chemical Abstracts Service has strong commitments in this area and has initiated policies and procedures which have proved effective. Nonetheless, research continues to improve and standardize ways of defining, processing, and storing data elements.

What seems to be the humblest of problems in data base management, that of providing for the data elements, may also be the most difficult to resolve -- the impacts of policies and procedures for the design and handling of data elements reach into file, program, and indeed overall system design.

1 7

6. References

[1] Anzelmo, F. D., A Data Storage Format for Information System Files, IEEE Trans. Comput. C-20(1) 39 (1971).

[2] Huffenberger, M. A., and Wigington, R. L., CAS Approach to Management of Large Data Bases, J. Chem. Inf. Comput. Sci. 15(32) (1975).

[3] The Data Base Administrator, Information Management Group, Information Systems Division, Guide, (Nov. 1972).

[4] Data Base Management System Requirements, Joint GUIDE-SHARE Data Base Requirement Groups, (11 Nov. 1970).

[5] Nerad, R. A., Data Administration as the Nerve Center of a Company's Computer Activity, Data Management, 11(10) 26 (1973).

[6] Requirements for the Data Dictionary/Directory Within the GUIDE/SHARE Data Base Management System Concept, Guide, (Nov. 1974).

[7] Uhrowczik, P. P., Data Dictionary/Directories, IBM Syst. J., 12(4) 332 (1973).

[8] Facility for Integrated Data Organization (FIDO) User Reference Manual, Chemical Abstracts Service, (April 1974).

[9] Management of Data Elements in Information Processing -- Proceedings of a Symposium Sponsored by ANSI and NBS, McEwen, H. E., Ed., (24-25 Jan. 1974).

[10] Technical Specifications for Standard File Format, Chemical Abstracts Service, (May 1973).

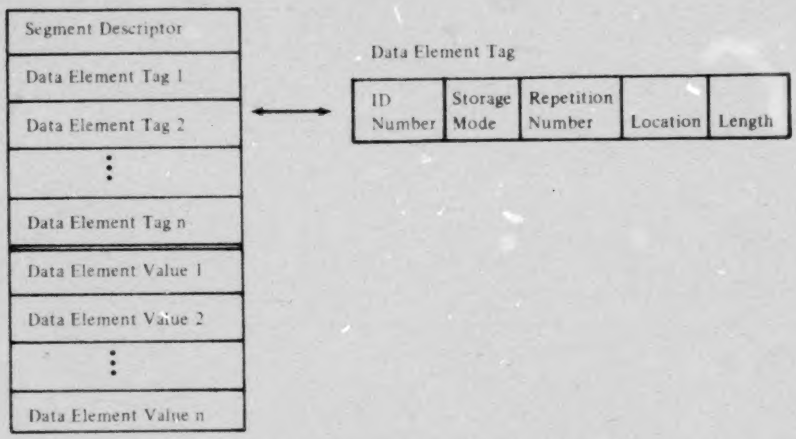

Segment Descriptor
Data Element Tag 1
Data Element Tag 2
Data Element Tag n
Data Element Value 1
Data Element Value 2
Data Element Value n
Data Element Tag
ID Number
Storage Mode
Repetition Number
Location
Length

Figure 1.

Schematic of a Standard File Format Segment. A typical file would contain thousands of segments like this.

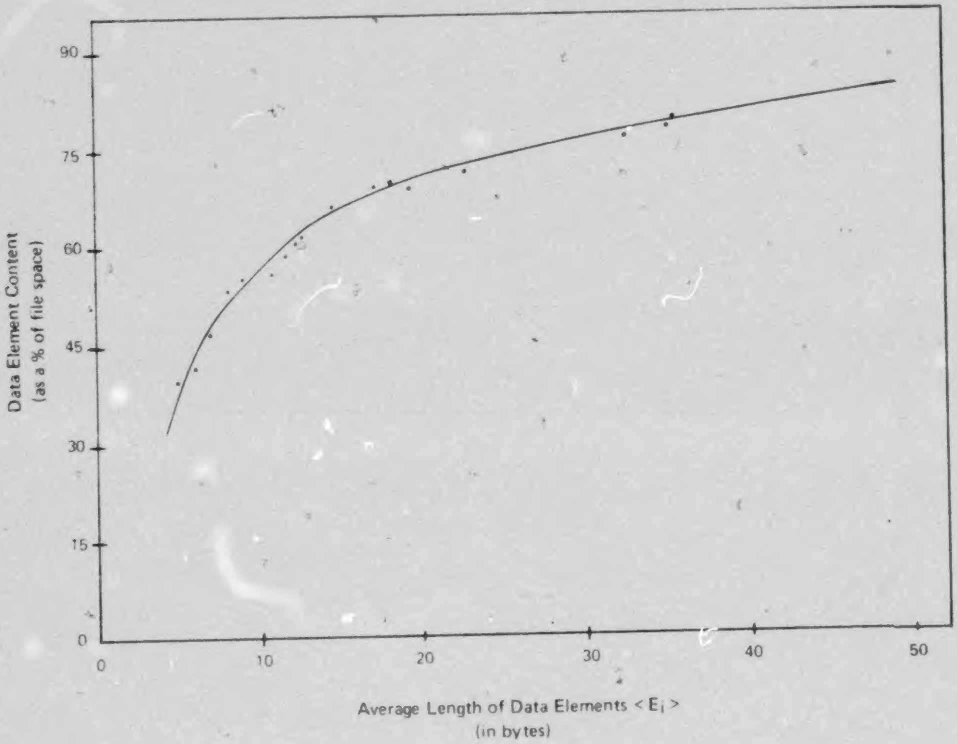


Figure 2.

Data Element Content Relative to Storage Requirements for SFF Files.

Addendum to
"The Design of Data Elements -- A Data Base Perspective"

M. A. Huffenberger

Chemical Abstracts Service
The Ohio State University
Columbus, Ohio 43210

Question: Is your system for data management available to other users?

Answer: The data management software we utilize was designed and programmed in-house. This system was designed specifically for the CHEMICAL ABSTRACTS processing environment, and if applied to other purposes, would require specific evaluation by the receiving organization. We are not in the software marketing business and accordingly, we do not distribute and support this system as a product. Nevertheless, a copy of the software and documentation can be made available for the cost of reproduction to organizations which have the background and are capable of integrating the system themselves without assistance from CAS for installation or trouble-shooting. The user reference manual for this system can be obtained from NTIS by ordering the "Facility for Integrated Data Organization (FIDO) User Reference Manual," May 1974, PB 236-020, and its first and second updates, PB 242-856, July 1975.

Question: Which Data Dictionary/Directory package are you using?

Answer: The DD/D package being used is under development by CAS.

Question: May we assume that your data base is so well organized that you can give a single command that will convert that funny DDMMYY form to the ANSI/ISO standard form for date - YYMMDD?

Answer: This question references a slide containing the description of the Publication Date data element, with the format DDMMYY. The ANSI/ISO standard should be adhered to for data interchange, but standardizing all date formats within an organization is a sensitive problem. While I discourage proliferation of date formats, there is some justification for multiple possibilities. In view of this, we do have a macro that performs conversions of date formats.

Question: In your quest for data independence, does your system put limits on the size of data elements? Does it handle the record level of definition at all?

Answer: As a practical matter, data elements can be any size. The upper limit is in the thousands of bytes. The system does handle the record level of organization. This is what enables this type of flexibility. Our stored records each contain a directory which identifies, describes, and points to the data element contents. Data elements can vary in length, appear in various storage modes, and even may be absent when they are optional. The resultant record structure is thus self-defining, and not at all the typical fixed format or format implicit in program code.

A Challenging Aspect of Word Processing

VICTOR G. KEHLER

Plans and Programs Division
Directorate of Administration
Headquarters United States Air Force
Washington DC 20330

Word processing utilizing computer assistance becomes a powerful tool to combat the inefficiencies of an information system that has historically been based on paper (hard copy) as the interchange media. The information under discussion is narrative data represented by correspondence, publications, messages (and this paper itself). This presentation traces the development, current status, problems and future of using computer assistance with word processing, computer output microform, electronic phototypesetters and soft copy terminals to process such information within the Air Force.

Key words: Communications; Computeroutputmicroform (COM); information composition; information disposition; information distribution; information reproduction; information retrieval; information storage; keyboards; microform; phototypesetting; publishing; text editing; typesetting; word processing.

1. The Air Force Version of Word Processing

Word Processing as a term was coined by industry. It conjures up different images to different people. In fact, it took many days of hard work for the Word Processing Ad Hoc Committee of The American National Standards Institute to develop the definition "word processing is the transformation of ideas and information into a readable form of communication through the management of procedures, equipment and personnel."

Within the Air Force the term used is Administration Centers. They are designed to fulfill the administrative needs of the customers they serve. These needs include telephone answering, filing, mail distribution, document copying, a source of office supplies, publication reference files and the making of reservations for scheduled trips in addition to keyboarding or typing.

Our word processing keyboards are not limited to those produced by one manufacturer. Although each center contains identical equipment of a single supplier, different centers will use different manufacturers equipment. Further, within a center the keyboards are modular in that some may be a single station, some dual, and some will have a communicating capability for computerized text processing. The two fold requirement is that the magnetic media produced on any machine within that center must be capable of being processed on any other machine in that center, and that at least one device in the center must be capable of communicating with a computer. The various centers, therefore, can interchange information by using the computer as the translation device. To put it simply, information recorded by Center A on a Brand X magnetic card device can be transmitted to a computer and recovered by Center B onto a Brand Y cassette.

Output from the centers can be either conventional hard copy correspondence, electrical messages in the optical character recognition font for scanning and electronic transmission via the Automatic Digital Network (AUTODIN), or a publication either on microfiche, on computer tape for electronic phototypesetting, or on hard copy for conventional typesetting.

2. Background

I work for the AF Director of Administration as a long range planner. I have been in this job since 1966. The Director of Administration is the paperwork manager for the Air Force. We do two things: we furnish policies and procedures to all Air Force personnel concerning paperwork; we also manage and operate certain services dealing with paperwork. We tell the people how to format letters, publications, studies and so forth; we tell them how to address things for distribution; we tell people how to set up their office files; and we tell the people when to get rid of those files and how to get rid of them. The services we provide include the official mail distribution which we pick up and deliver Air Force-wide; we provide the reproduction function Air Force-wide, by operating printing and reproduction facilities; we provide the distribution facilities to take care of the storage and distribution of all our departmental publications; we provide microform service centers to assist offices in converting files to microfilm and we also provide the records staging areas for short term retention of records outside of the normal office filing cabinet.

The long range planning team which I head up was initially established to facilitate the conversion of the publishing process from conventional "buy it up town" typesetting to an in-house electronic or computerized phototypesetting system. The system was designed to reduce publication cost. The phototypesetting system which is the Mergenthaler Linotron is installed at HQ Air Force Logistics Command at Wright-Patterson AFB, Dayton, OH. It accepts information contained on computer tapes and also accepts graphics which are converted to videotape and then marries the two via computer to produce a typeset page of information containing both narrative style information and graphics. The system was acquired to accomodate statistical supply type catalogs. The cost trade-offs were the twenty to forty-five dollars a page being paid to commercial typesetters versus approximately one to three dollars for computerized typesetting.

The Linotron system has worked and is working quite well for typesetting that information that is contained in a computer. Immediately back in the late 60's we took a hard look at our total information file which consisted primarily of publications and correspondence. Our idea was to expand use of the Linotron to include such narrative information. However, at that time, about the only equipment available was the old work horse International Business Machines magnetic type selectric typewriter that could with a little work be translated into a computer. The biggest plus was that we had a captive audience in that we controlled all of the typists who were generating this information that we wanted printed. We did see that if we could break down the wall between the publishing system and the office environment we could capture the information at its time of initial keyboarding and then merely massage it by our publication people before sending it to the computerized publications system. However, the system we were looking at kept getting bigger in its scope and we ended up advocating a 4C Keyboard; one keyboard that would prepare correspondence, would prepare information for composition, would prepare communications such telegrams for electric transmission, and finally would prepare computer data for entry into the conventional bookkeeping type computerized application such as accounting and finance systems, supply systems and so forth. The bond of commonality was that the office environment which is present in any mission element within the Air Force had a typist, a typewriter, a filing cabinet, and an in and out box. The staff officers or project officers who use the services of this clerk-typist or clerk-steno were the people who were initiating correspondence publications, messages and computer data.

And so we had shortly developed on paper an information processing system which could reduce our operating costs,increase the speed and accuracy of information handling and also reduce the usage of paper. At this point we were actually using such equipment and demonstrating that it could be done.

Very simply to describe the system: the requirements are that all information at the time it is initially keyboarded will be recorded on a magnetic media for future playback, revision or tranmmission to a computer. Publications and larger studies would be transmitted to a computer for the more extensive text editing capabilities. Coordination of documents could be handled with keyboards and associated cathode-ray tubes within the staff officer area for information retrieval, display, and minor editing functions. The publications could be electronically reviewed by the publishing experts and the output of the system could be soft-copy such as display terminals or hard-copy such as microfiche, microform or paper. The two largest exceptions to the information susceptable to this new form of processing are classified information or information prepared for transmission to other than Air Force activities.

Thus the results of our studies were to advocate the use of the computer for narrative information processing which means the creation and editing of information its coordination its distribution its storage and recall and finally its disposition.

3. Word Processing, Computer, Microform Interplay

Now lets make the relationship between our publishing system, the use of computer output microfilm equipment and the word processing concept. The historic paper publishing system within the Air Force said that all offices who wanted a publication would provide typewritten manuscript copy to the publishing people. The manuscript copy was edited then sent up town for typesetting. After typesetting it was run through the printing plant presses and the publication was eventually ready for distribution Air Force-wide. With the advent of word processing equipment it meant that all of the typists could now produce manuscript copy much faster. But by taking another step and giving the communications capability to the word processing equipment we can now dump the information into a computer where it can be coded by our publishing people for output either to an electronic phototype-setter such as the Linotron at AFLC or to a computer output microfilm equipment for production on microfiche. We are now very active in converting our publications from paper to microfiche. We are using the Department of Defense standardized reduction ratios of either 24X or 48X. We also have standard viewing equipment or microfiche readers available within the Department of Defense which have both 24 and 48X lenses so that the choice of the reduction ratio presents no problem to the user. We are using 24X to direct film the camera-ready copy being prepared by word processing equipment or standard electric typewriter which for some reason or another cannot be computer processed. We are using 48X for those publications which are computer maintained and can be produced by computer output microfilm equipment. The biggest hangup preventing us from an accelerated conversion of publications to microfiche is the lack of sufficient viewing equipment Air Force-wide to read the publications in microform.

We have two big problems when using the computer and COM to produce narrative style publications. The first of these and the most critical is that many of our publications contain graphic information or line drawings. We need a system that will produce simple graphics at a cheap price on microfiche. The second problem is that we lack a good automated package or software program to automatically generate the index of the publication so that the viewer can glance at the table of contents and see the XY reference for the particular paragraph that he wants to read.

We are now engaged in producing some of the publications on microfiche both at the 24 and 48X reduction ratios. They include Air Force Manual 300-6 concerning management of a data processing installation, Air Force Regulation 0-12 an index to forms, and many of the computer operator manuals which give instructions to computer operators and the functional user support manuals which are the manuals given to the customer on the data automated system explaining to him how to feed the system and what he can get out of it. In addition to the narrative we are also involved with the use of COM to replace the conventional computer output reports, the statistical reports at many of our Air Forces locations. Because of the speed of COM we are using a single COM per installation to produce both narrative and statistical type information thus marrying the administrative and the data automation or computer world.

To give you a brief summary of the relationship of word processing equipment and micropublication we now have many manuals and regulations in the computer for text processing and electronic maintenance. Some of these publications are phototype set, some are produced by COM, some are produced on microfiche from direct film procedure of the camera-ready copy, some are rekeyboarded or typeset, and some are printed "as is" or camera-ready typewritten copy.

So now in addition to word processing equipment we have administration involved in computer output microfilm and in software design for text processing and information storage and recall processed by the computer. With the addition of communication capability to the word processing equipment we have completed the loop for tying together information creation, coordination, transmission, reproduction and distribution.

4. The Privacy Act of 1974 and Word Processing

In May of 1975 the Air Force, along with the entire Department of Defense, was tasked to prepare on computer tape and in upper and lower case characters narrative information concerning all records systems we maintained that contained an individuals name, and we had to have the tape ready by mid-June 1975. We gave instructions to the field to furnish us the information on magnetic media prepared by word processing equipment in upper and lower case. If they could not do so, we agreed to accept upper case only computer tape generated by the punched card approach. Preparation instructions for either the word processing or punched card approach were identical with but one small difference: We used a unique 13-digit code as an identifier, and this code had to be in columns 1 to 13 of each and every punched card, but was only required in the first line o media prepared on word processing equipment. The reason - each punched card contained but one line of 80 characters, while the word processing media had many lines so we didn't worry about dropped or lost cards.

The word processing media were dumped into a time share system via our word processing communicating devices. The computer tapes were fed into the same computer. We then used the word processing equipment on line to edit and massage the raw data for preparation of the computer tape required by the Department of Defense.

About the upper case only on those records prepared via punched cards. We did an automatic case reversal changing all such data to lower case only with the exception of the first 13 characters (the unique identification number of each system). We used global commands to change common terms as united states air force into United States Air Force. A global command is one instruction that says look in the entire file for this term and every time you find it, this is what you do.

The final output, about 2,000 pages, was completed and on time thanks to the computer, word processing and the rapport we have with our data automators.

5. The Future

Getting from a system on paper to an operational system does present a few problems. We have two approaches:

a. We have established a project to look at the Air Forces information needs to design a system to satisfy these needs. The big hang-up here is the immediate and emotional reaction when the people realize we are saying hey we have a system that doesn't need paper. You in the audience by now recognize this as the wired city concept. Approach A, therefore says we have to define the job that we want to do and then we design the system to do that job.

b. Approach B is forced on us by the reality of what is happening today. There are many software packages that give you text editing capability and information retrieval capability. There are many different phototypesetters and computer output microfilm equipment on the market. There are many different companies putting our word processing equipment using tapes, cassetts, cards, discs or mini-computers and some of these use the cathode-ray tube while others use communication facilities. Within the Air Force and its a big organization we have many commands and many bases. Equipment salesmen are out there now getting contracts on different bits and pieces of the system. We do not have a central control to the extent where one office and one office alone contracts for this equipment for use within the Air Force. What we are doing is alerting and advising our people as to what the system can do if built correctly and trying to keep them from painting themselves into a corner by getting equipment that cannot mesh with the total information processing requirements.

6. The Challenge

The challenge to the technical computer people is that we administrators who set policy and help people run the information system know pretty well what we want to do. We need the systems and the software to handle narrative information (with graphics) and with a very simple man machine interface. We need it simple because who will be on the other side of the keyboard, it will be our clerk stenos our clerk typists, our staff officers, and eventually higher level management people. We should be able to create information using conventional typing techniques and with conventional touch keyboards. I am very much aware that we do need a forty-eight keyboard to accommodate the American Standard Code for Information Interchange, the ASCII code. I also happen to be the Department of Defense representative to the American National Standards Institute Committee X4 which is responsible for developing the standard keyboard. My position is that we do need the forty-eight keyboard but we cannot disrupt the touch keyboard that all of our people are now used to. This means that the colon, semicolon, the double comma, the double period, should remain where they are. This creates some problems in the standard keyboard to accomodate the ninety-some graphics characters of ASCII. But it can be done.

The Software developed by the programming specialist must be simple. We need common sense type instructions to format and rearrange information for information processing, for interfacing with various phototypesetters and computer output microfilm. We need automatic indexing and insofar as possible automatic function code insertion to drive output devices. We need reliable information retrieval software, we will want to be able to find related information and we will also want to automatically dump out information which is eligible for disposition. Again since we are with the government we must meet the National Archives Records Service criteria on the form of media for material being retired.

And now I will draw a conclusion, not a summary, but a conclusion. The first challenge is in changing people from their established methods of using paper for communication and to get their job done. Word processing equipment married to a computer gives us the mechanics to build the system to wean people away from paper, but in building it we must keep the users needs at the forefront. The system will have to be simple, reliable, simple, less costly, simple, efficient, and simple.

Data Element Lexicon Needs A New Home

Richard J. Kirkbride
Computer Specialist

National Military Command System Support Center
Military Studies and Analysis Directorate
Logistics Data Division
Washington, D. C. 20301

A large scale Data Element Lexicon (DELEX) System exists for operation on IBM 360. The coding is largely FORTRAN with ALC for I/O. The FORTRAN is used to obtain benefit of algebraic compare instructions to minimize sorts and maximize indexing efficiency.

The system was developed in 1970-72 and made operational by Computer Sciences Corporation under contract to Defense Communications Agency (DCA). The DCA has completed the project and no longer maintains or operates DELEX. Any government organization interested in further utilization and maintenance of the DELEX can obtain a tape copy of the DELEX in five reels from:

DCA/NMCSSC
ATTN: Technical Director
Pentagon, Rm. MF627
Washington, D. C. 20301

CHECK CHARACTERS AND THE "SELF-CHECKING STRING" --
What, Where, Why, When and How

J. R. Kraska, J. R. Nelson

The Upjohn Company
Kalamazoo, Michigan 49001

and

E. Hellerman

Bureau of the Census
Suitland, Maryland 20233

"Self-checking strings" are character strings, used as data-base keys, which must pass a self-validation check before the keyed information can be transferred. Basic to self-checking strings is the concept of "check characters".

The criteria for construction of a self-checking string are given. To enable the potential user to decide what is best for him, a structure for cost/benefit analysis is outlined. The identification of a predominant class of systems used to compute and validate a self-checking string is made. Unified and detailed procedures are given for 1) defining a self-checking string, 2) computing a valid self-checking string, and 3) validating a string. Many of the systems in use today are summarized within the framework of this unified classification scheme. Contrasts are made with respect to power (the ability to detect errors). Examples are given of string definition, computation, and validation for two systems which combine the best of power and efficiency. FORTRAN code is used to illustrate the ease of implementation for these two systems.

Key words: Check characters; classification; cost/benefit; data key validation; efficiency; FORTRAN; information transfer; key error types; power; self-checking string composition; software.

1. Introduction

One problem - actually a major problem - in the transfer of data from user input documents to computerized storage (see fig. 1) is "routing" the right information to the wrong place. Valid transfer depends fundamentally on the transmission of a valid key to the software procedure doing the "routing" (for examples of commonly used keys, see fig. 2). What makes a key invalid? The answer is simply that there is an error in the characters in the key. These errors can be introduced by the user in the preparation of his input documents and/or by the system data handlers, whether they are human (e.g., a keypuncher), or mechanical (e.g., an optical character reader). Figure 3 summarizes the types of errors that are frequently made [1,2][1]. No comment is to be made about the likelihood of each type of error.

---

[1] Figures in brackets indicate the literature references at the end of this paper.

It would seem that having a procedure which checks the key for validity would be a good idea; and it would be important to have a procedure which has a low probability of validating an invalid key.

An approach frequently taken is to imbed one or more characters called "check characters" in the key. This new string is called a "self-checking string". One or more compatible check procedures can then be used, both to compute valid check characters and to validate transmitted keys prior to routing the information to which the keys are attached.

In the remainder of this paper, we shall define the composition of the self-checking string, define check procedures, and characterize the major class of check procedures found in the literature. The earliest occurrence which we are aware of to a specific procedure which is a member of this class is due to IBM [3]. We shall define the general computation and validation steps which allow implementation of any of the procedures within this class. We shall provide a framework for the analysis necessary to decide whether a check character system is cost-beneficial for a particular application. Finally, we shall present criteria for deciding which check procedure(s) to use, and introduce two new powerful and efficient algorithms.

2. Composition of a Self-Checking String (SCS)

In general, a self-checking string is composed of user pre-assigned characters, arbitrary characters, and check character(s). Let the number of pre-assigned characters be denoted by n_p, arbitrary characters by n_a, and check characters by n_c. The total string length l is thus equal to $n_p + n_a + n_c$. The order of the characters is arbitrary. However, for "convenience" the check character(s) are usually positioned at the end of the string.

Example: With $n_p = 3$, $n_a = 4$, and $n_c = 2$, one possible ordering is

pppcaaaca

where "*p*" denotes a pre-assigned character, "*a*" an arbitrary character, and "*c*" a check character.

Each value of the pre-assigned characters *ppp* represents a unique "class" of importance to the user. The characters *aaaa* are used to uniquely identify members within each "class". However, depending on the procedure used, certain combinations of these characters may not yield valid strings, and must be discarded (see app. 1). Thus the number of distinct valid combinations of the n_a arbitrary characters must be large enough to specify each member of the "class" having the maximum number of members. This is the criterion used to decide the number of arbitrary characters n_a.

The number of check characters n_c depends on the particular "check procedure" chosen.

3. Definition of a Check Procedure

A check procedure is a set of steps for:

1) computation of a valid SCS of specified composition and
2) validation of strings which were ostensibly computed as above.

129

4. Characterization of the Major Class of Check Procedures

The literature makes reference to a large number of check procedures. Some are applicable to digit strings only [1] and others to general alphameric strings [4]. Although the characteristics of these check procedures seem at first glance to be widely variant, we have come to realize that in fact most of the procedures have identical characteristics. We have developed the following expression which identifies these characteristics (for the definitions of terms, see app. 2):

$$\{F, I, O, P, M, C, D, T\}$$

The eight elements (characteristics) in this expression are:

1. F-the function which generates the weight string W, one value for each character position of the character string.

   F may be specified in a variety of ways and can generally take the form of a sequence of numbers of total or partial string length l. Some examples are:

   a) (1,a) repeated sequence extended to the length of character string; "a" is a digit, usually 2 or 3.
   b) (a,b,c) repeated sequence of digits extended to the length of the character string; the characters denote arbitrary digits.
   c) geometric progression (e.g., powers of 2).
   d) prime numbers.
   e) powers of 10 where the exponent of 10 is the "index" defined below.
   f) any function that can supply a weight value to every position of the character string.

2. I-the "index base" - either 0 or 1, which is the value of the first index position in the string. The symbol "i" is used to denote index position in the following material.

3. O-string "orientation". The character "R" is used to specify right to left orientation. The character "L" is used to specify left to right orientation.

4. P-the "products digits transformation".

   The "products digits transformation" strategy can be expressed as one of the following:

   a) Whole Products Summed (WPS) - the products of the string character numeric values and their associated weights are summed.
   b) Products Digits Summed (PDS) - the digits from the products of string character numeric values and their associated weights are summed.

5. M-a modulus. Thus M = 11 indicates that a remainder modulus 11 is computed.

6. C-the number of check characters to be evaluated.

7. D-the value which is checked against the result of the validation procedure to determine if the self-checking string is valid.

8. T-the table of character values (for the case of numeric strings, the table is the identity transformation).

Since the characteristics above are identical for all members of the "class", the sequence of steps for computation and validation are also identical. Common to computation and validation is a sequence of steps which we call the core computation (see app. 3). Steps necessary for computation (step 1 of the "Definition of a Check Procedure") are outlined in appendix 4, and steps necessary for validation (step 2 of the definition) are outlined in appendix 5. The characteristic values for selected check procedures belonging to this class are listed in appendix 1. An example of computation and validation for a specific check procedure (#10 of app. 1) is included in appendices 4 and 5.

5. Costs and Benefits

In considering the economic impact of a checking system there is one simple specific question to answer:

> Is the cost of development and implementation of the check system less than the cost of correcting the errors that would be made if the check system were not there?

Frequently the answer to this question is yes. The errors introduced into complex data base systems tend to propagate. One system support analyst at The Upjohn Company has said, "90% of the errors people make while attempting to use our computerized systems result from attempts to correct other errors." If this is generally true, one mistransfer, which most likely would have been prevented by a good check procedure, creates nine additional errors before it is corrected.

Furthermore, the operational cost of an efficient check system is almost negligible compared to the cost of inputting the data; and the cost of inputting the data is negligible compared to the cost of error correction.

The following is an outline for a quantitative cost-benefit analysis:

1. Estimate the total development cost for the checking system (We suggest less than $2000).
2. Estimate the operating cost for the checking system per string. (Test data for one "good" procedure suggests that one penny per 1000 strings is a reasonable estimate.)
3. Estimate the volume of strings to be processed per time period.
4. Estimate the fraction of strings processed which will be in error.
5. Estimate the net benefit for each error detected - i.e., the dollar value of cost to correct the error if undetected less the dollar value of cost to re-process after detection and correction.

As an example, consider a hypothetical system processing 2000 strings per month, with an error frequency of 0.5%, and a net benefit of $20 per detected error. Then assuming that a check procedure which catches ~100% of the most likely errors is being implemented, our hypothetical check system pays for itself in 10 months.

It is very important to realize that the development cost is independent of the size of the system, and the operational costs are effectively zero. Therefore, the benefit to cost ratio is directly proportional to the product of the volume of strings transmitted over the life of the system, and the net benefit of detecting/correcting errors -- a remarkable result.

6. Choice Criteria

We have strongly advocated using check procedures but have said nothing about how to decide which procedure(s) to use. Some are better than others! The decision should be based on the following criteria:

1) Checking Environment (hardware or software).

2) Power - for each type of error (see fig. 3) the percentage that will be detected (see app. 1).

3) Operational Efficiency - the operating cost for the validation of one string.

4) Inconvenience

   a. of extra string length due to check characters,
   b. of discarding strings for which no valid check character(s) exist.

The most important of the above criteria is the environment for the check system, as it can severely limit the choice of check procedure(s) to use.

Initially, check systems were implemented exclusively on hardware keying devices [3,5]. Because of inherent limitations (e.g., inability to do complex integer division) there were only a few procedures which could be implemented. This is still true today. These procedures have become widely known and widely used: (cf app. 1, #5, #8, #10).

Now, however, something different is happening. Because of the speed of central processors in modern computers it has become economically feasible to implement check systems as software pre-processors. To date, it has been the practice to implement in the software environment the hardware oriented procedures noted above. This is unfortunate, since many new procedures have been devised which cannot be implemented on hardware, but which are much more powerful, more efficient, and easier to implement on software than the hardware oriented procedures (cf app. 1, #1, #2, and app. 6).

We envision the following type of implementation:

> The check system is a software program positioned between the input software and the "routing" software. As a string is read it is checked, and either the associated data is routed, or an entry is made in an error report. Errors are corrected and re-processed all at once. This is in contrast to the hardware oriented procedure which catches errors at the time of keying.

For a somewhat more detailed discussion of implementation alternatives see Mason and Connelly [9].

We believe that check systems implemented in a software environment rather than at the earlier keying stage are less costly to develop and are more powerful and efficient in operation. As an example, a particular alternative to developing one software program to do checking might be to modify six keying machines by adding check feature modules. This latter alternative is much more costly.

Once the choice of the "environment" is made, the remaining criteria are considered. The user would like to maximize "power" and "efficiency" but to do so he is usually forced to accept the unwanted side-effect of "inconvenience". The best approach for the user is to establish a maximum level of inconvenience which is acceptable and then choose the best of all procedures which do not exceed this level.

7. Implementing a Check System

To implement a check procedure in appendix 1 do the following:

1) define simultaneously the composition of the self-checking string and a choice of check procedure(s).

2) follow the generalized steps of appendices 3-5 for computation of valid strings and validation of strings.

There are a number of important considerations concerning implementation:

- Our scheme for characterizing this "class" of check procedures was meant for purposes of generalization, unity, and clarification. However, in some instances we have inadvertently compromised implementation efficiency. The user is encouraged to recognize that characteristics of specific procedures make it possible to either

a) define alternate steps for computation/validation and/or,
b) delete steps.

- Not all check procedures are compatible with an arbitrarily specified SCS. Compatibility depends on

  a) the number of check characters specified in the SCS composition,
  b) the type of check character (digit or alphameric).

- Multiple check procedures may be defined for a given SCS. Two examples:

  a) Two procedures are defined for an SCS having a single check character [6]. For a string to be valid, both procedures must validate.
  b) Two procedures are defined for an SCS having two check characters [4]. One procedure computes and validates using one character, the second computes and validates using the other character. Both must validate for the string to be valid.

8. Technical and Experimental Results

During a search to find check procedures which would be very efficient to implement in a software environment, while (hopefully) sacrificing little in terms of power, a startling and (we think) important discovery was made. Stated simply, the discovery is this: For numeric strings, dividing the entire string, treated as one number, by another number, known as the "Modulus", and taking the remainder, implicitly defines all but one (D, the digit to be checked against) of the "characteristics" of a check procedure.

The advantages of such a check procedure should be obvious. Integer division is many times faster than considering the individual digits of a string and multiplying by "weights".

The only unsolved problem was to find a modulus (or moduli) which defined a check procedure powerful enough to consider using. It can be shown that satisfying the following criteria would be sufficient to provide a check procedure at least as powerful as any other in the "class":

1) The modulus must be prime.

2) To catch single substitution errors, the modulus must be greater than the maximum number of distinct characters possible in one position - i.e., 10 [7].

3) To catch all transposition errors (both adjacent and non-adjacent), the implicit effective weights (see [1] p 86) must be distinct.

Important Note: Any modulus M satisfying these criteria will catch all single substitution and all transposition errors. Furthermore, it will catch the fraction 1 - 1/M of the combination errors.

The smallest modulus satisfying these criteria is 17. The implicit effective weights for 17 repeat in groups of 16, so no transposition errors are possible in strings of less than 17 characters. Note that in the computation of valid strings, if only one check digit is used, 7/17 (about 41%) of all strings must be "discarded". If two check digits are used, a much better modulus to choose is 97 (It is no coincidence that Whiting's [8] base 2 check modulus is 1100001 which is equal to 97 base 10.) as the number of combination errors not caught is cut by a factor of about 6, and no strings are discarded.

In an attempt to measure the increased efficiency of these two new check procedures, MOD 17 and MOD 97, tests coded in FORTRAN were run on our IBM 370/158 VS-2 computer facility. The tests indicated that these procedures executed at least 4-6 times faster than two other procedures tested (Procedures #3, #8 in app. 1). Based on internal user rates, the cost per validation was less than one penny per thousand.

We think that the discovery of these two new procedures, easy to implement and fast to execute on digital computers, and more powerful than other procedures discovered to date, is significant.

Appendix 1

#	M	P	O	I	C	D	F①	% ERRORS CAUGHT SS	AT	NAT	C	Discard Fraction	Comment⑦
1	17	WPS	R	0	1	0	10^i	100	100	100	94.2	.41	Less than 17 characters.
2	97	WPS	R	0	2	0	10^i	100	100	100	~99	0	
3	11	PDS	R	-	1	0	(1 2)	100	97.8	<50②	90.9	.09	
4 [1]	11	WPS	R	0	1	0	2^i	100	100	100	90.9	.09	
5 [1]	7	WPS	R	0	1	0	10^i	93,3	93.3	93.3	86,4	0	Less than 7 characters.
6 [1]	11	WPS	R	0	1	7	10^i	100	100	<50②	90,9	.09	
7 [1]	11	WPS	R	-	1	0	1,3,7,13...③	100	100	~97⑥	90,9	,09	
8 [1]	10	WPS	L	-	1	0	(1,3,7)	100	88.9	~60②	90	0	
9 [1]	11	WPS	R	-	1	0	(1,2,3,4,5,6,7,8,9,10)	100	100	100	90,9	.09	Less than 11 characters.
10 [3]	10	PDS	R	-	1	0	(2 1)	100	97.8	<50②	90	0	
11	10	PDS	R	-	1	0	(1,2,3,4,5,6,7,8,9)	100	97,8	~85	90	0	Less than 10 characters.
12	23	WPS	R	0	1④	④	10^i	100	100	100	95.7	0	See note④.
14	23	WPS	R	0	2⑤	⑤	10^i	100	100	100	99.8	0	See notes④ and ⑤.

①For orientation L, the weight string is applied in the order that it appears starting at the left-most character of the string, For orientation R, the weight string is applied in the reverse order of appearance starting at the right-most character of the string. IMPORTANT NOTE: The orientation L/R is irrelevant as far as power and efficiency is concerned.

②Dependent on length of string.

③Successive primes, excluding 5 and 11.

④Check character at end of string is alphabetic (from table T of 23 alphabetic characters) - result of core calculation is checked against this character.

⑤Two alphabetic check characters - divide string by 23, and then its quotient by 23.

⑥For strings of length $l \leq 10$.

⑦Restrictions on string length are made because additional non-adjacent transposition (NAT) errors will remain undetected in longer strings.

Appendix 2

Definitions:

1. **Character String** - a sequence of alphabetic and/or numeric characters. This may also be referred to as a "string".

2. **Digit String** - a character string composed of numeric characters only.

3. **Character Value** - the numeric value associated with an alphabetic or decimal digit character. Normally, a decimal digit represents its own value.

4. **String Length** - the number of characters in the string including the check character.

5. **Weight String** - a sequence of values in one to one correspondence with the character string positions which act as weights (multipliers) to the corresponding string character values.

6. **Weight Function** - the rule by which the weight string is generated.

7. **Orientation** - the direction in which the string is operated upon - i.e., right to left or left to right.

8. **Index** - the position of a character in a string with respect to orientation.

9. **Index-Base** - details whether the first index is labelled zero or one.

10. **Modulus** - a number used as a divisor so as to produce a remainder between 0 and Modulus minus 1 inclusive.

11. **Transcription Error** - the substitution of one or more incorrect characters for correct characters in a string.

12. **Substitution Error** - same as transcription error.

13. **Transposition Error** - the entry of the correct character for the ith position into the jth position and the character for the jth position into the ith position.

14. **Combination Errors** - a combination of transcription and transposition errors (addition or omission of characters is also considered a combination error).

15. **Products Digits Transformation** - a mapping of the products of the elements of the weight string and the corresponding elements of character values.

16. **Check Procedure** - a particular algorithm for computing and validating self-checking character strings.

Appendix 3

Core Computation

Steps:

1. Decompose the string into its characters.

2. Replace each character with the appropriate character value (only necessary for strings with non-numeric characters).

3. Multiply each element of the result in step 2 by the corresponding element of the weight string W.

4. Make the "products digits transformation" P.

5. Evaluate the remainder of the sum in step 4 modulus M. Call it r.

Notes: 1. The core computation for all members of the class of check procedures that we are considering can be performed as outlined above. However, for any particular procedure there may be an alternative sequence of steps which is more efficient to implement and/or execute. Specifically, see Appendix 6 for the MOD 17 and MOD 97 check procedures, for which the core computation reduces to one step (step 5 above).

Example using check procedure #10 from Appendix 1:

Given the string

3141593

1) 3 1 4 1 5 9 3
2) not necessary
3) 3 1 4 1 5 9 3
   x x x x x x x
   2 1 2 1 2 1 2
   = = = = = = =
   6 1 8 1 10 9 6
4) The transformation:
   6 + 1 + 8 + 1 + 1 + 0 + 9 + 6 = 32
5) $32 \equiv 2$ MOD 10. Thus r = 2.

Appendix 4

Computation of a Valid Self-Checking String

Steps:

1. Establish the "composition" of the string and choose the check procedure to be used (see app. 1). Note: the number of check characters n_c must be consistent with the check procedure.

2. Construct a table of length M (index origin 0) according to the following steps:

   a) Initialize each table entry to a code denoting an invalid string.
   b) Substitute zero (0) for all "*p*" and "*a*" characters in the string. Call the result string S.
   c) For each possible combination of check character(s)

      i) substitute the character(s) into S in the "*c*" character positions.
      ii) perform the "core" computation (see app. 3) on S to obtain the remainder r.
      iii) subtract r from M + D and take the remainder modulus M. Call it r'. Use r' as the table index to store these characters.

   Note: For any particular check procedure this table need only be constructed once.

3. Specify the n_p - "*p*" characters for the string. Assign arbitrarily the n_a - "*a*" characters (possibly according to a sequential rule), and insert zeros for the "*c*" character positions.

4. Make the "core" computation to obtain r.

5. Use r as an index to the table.

6. If r points to valid check character(s), insert these in the appropriate positions in the string.
7. If r points to an invalid string code, discard the string.

Example:

1. String composition:

   pppcaa

   ($n_p = 3$, $n_a = 2$, $n_c = 1$, $\ell = 6$)

   Check procedure: #10, Appendix 1.

2. a) Table

index	check character
0	i
1	i
2	i
3	i
4	i
5	i
6	i
7	i
8	i
9	i

The initialized table;
"i" denotes invalid string.

b) S is 000·00

c) i-iii)

Table

S	r	r'	index	check character
000000	0	0	0	0
000100	2	8	1	9
000200	4	6	2	4
000300	6	4	3	8
000400	8	2	4	3
000500	1	9	5	7
000600	3	7	6	2
000700	5	5	7	6
000800	7	3	8	1
000900	9	1	9	5

Note: the possible check characters are the integers 0-9 inclusive.

3. A particular string of the above composition is 321099 (zero substituted in check character position).

4. The core steps are:

```
1)  3  2  1  0  9  9
2)  not necessary
3)  3  2  1  0  9  9
    x  x  x  x  x  x
    1  2  1  2  1  2
    =  =  =  =  =  =
    3  4  1  0  9 18
4)  The transformation PDS:
    3 + 4 + 1 + 0 + 9 + 1 + 8 = 26
5)  26 ≡ 6 MOD 10.  Thus r,= 6.
```

5. The sixth position in the table points to valid check character 2.

6. The self-checking string is

321299

Appendix 5

Validation of a String

Steps:

1. Make the "core" calculation to obtain the remainder r.

2. If r is equal to D, the string is valid; otherwise, it is invalid.

Example (continuation of example in Appendix 4 which uses check procedure #10 from app. 1):

Introduce a substitution error in the valid self-checking string computed in Appendix 4.

$$\underline{3}21299 \xrightarrow[\text{error}]{\text{substitution}} \underline{8}21299$$

1. The core steps:

1) 8 2 1 2 9 9
2) not necessary
3) 8 2 1 2 9 9
x x x x x x
1 2 1 2 1 2
= = = = = =
8 4 1 4 9 18
4) The transformation PDS:
8 + 4 + 1 + 4 + 9 + 1 + 8 = 35
5) 35 ≡ 5 MOD 10. Thus r = 5

2. Since $r = 5 \neq D = 0$, the string is not valid; an error has been detected.

FORTRAN Code for MOD 97 and MOD 17 Procedures

Let N be a particular string of given composition (i.e., a string specified by following steps 1. and 3. of app. 4) restricted as follows:

1. N is numeric.
2. The check digit(s) in N are at the end of the string (e.g., *ppaacc*).

FORTRAN code for "computation" of a valid self-checking string L is:

Computation MOD 97:

```
   L = N/97 * 97 + 97
```

Computation MOD 17:

```
   L = N - (N-1)/17 * 17
   IF (8.GT.L) GO TO 10
   L = 17 + N - L
   "valid" processing
          .
          .
10 "invalid" processing
          .
          .
          .
```

FORTRAN code for "validation" of a string LP ostensibly computed as above is:

Validation MOD 17 or 97:

```
   IF (LP.NE. LP/M * M) GO TO 10
   "valid" processing
          .
          .
          .
10 "invalid" processing
```

where M is either 17 or 97.

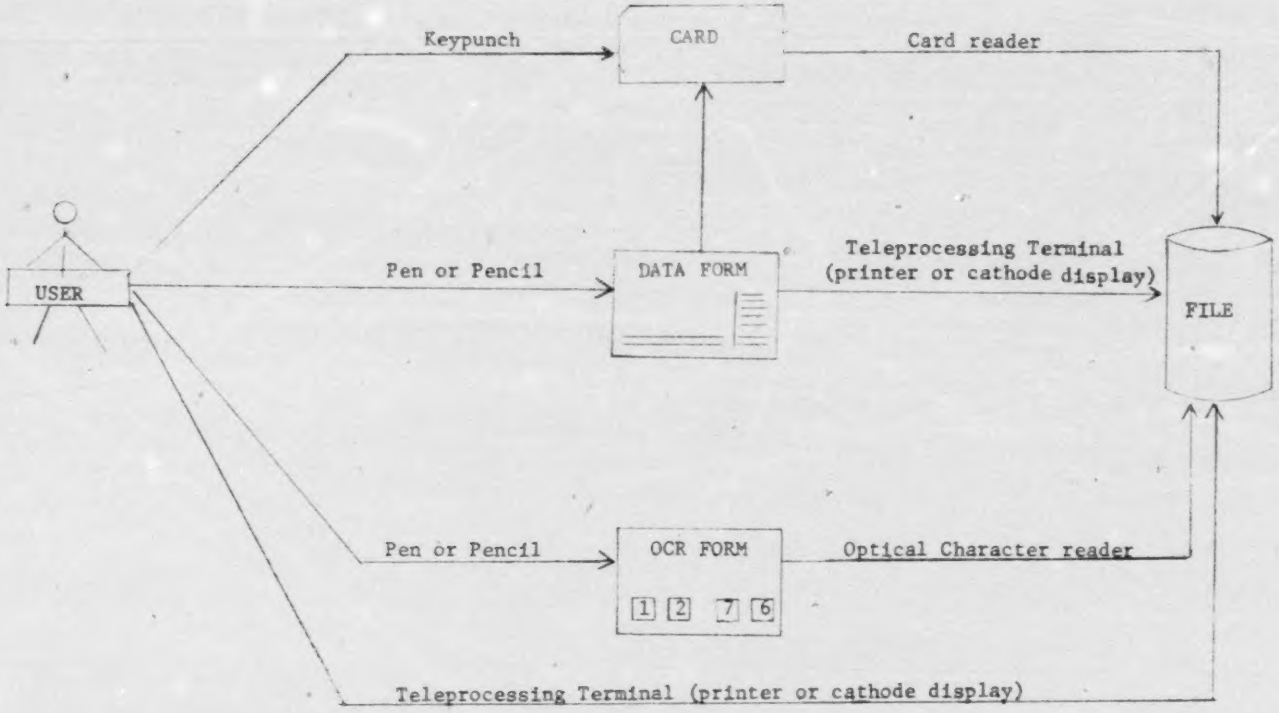

Keypunch
CARD
Card reader
USER
Pen or Pencil
DATA FORM
Teleprocessing Terminal (printer or cathode display)
FILE
Pen or Pencil
OCR FORM
1 2 7 6
Optical Character reader
Teleprocessing Terminal (printer or cathode display)

Figure 1.

TRANSFER OF INFORMATION FROM USER TO COMPUTERIZED DATA FILE

Account No.	Bank Deposits Credit Cards
Item No.	Grocery Items Lot No. Automobile Parts Inventory
Customer No.	Airline Passenger Laundry Ticket Freight Container
Registration No.	Social Security Dog License Automobile Registration
Survey No.	Harris Poll Government Census

Figure 2.

EXAMPLES OF COMMONLY USED KEYS

VALID STRING	ERROR TYPE	INVALID STRING
321299	SS	821299
321299	DS	821209
321299	AT	312299
321299	NAT	391229
321299	C	871200

SS = Single Substitution

DS = Double Substitution

AT = Adjacent Transposition

NAT = Non-adjacent transposition

C = Combination (random)

Figure 3.

EXAMPLES OF ERROR TYPES

References

[1] Herr, J. R., Self-Checking Number Systems, Computer Design, p 85-91 (June 1974).

[2] Radack, B., Check Digit, IRS Memorandum to E. Hellerman (Feb. 4, 1975).

[3] _________________, IBM 24-26 Card Punch Bulletin Self-Checking Number Feature, G24-1057-0 (1952).

[4] Jeane, M. D., Self-check Number Devices -- Check Digit Systems, U. S. Dept. of Commerce Memorandum to S. Dolleck (Jan. 9, 1970).

[5] _________________, IBM 24 and 26 Card Punch Bulletin Self-Checking Number Feature Modulus 11, and its Associated Self-Checking Number Generator, Modulus 11, G24-1022-1 (1960).

[6] Anderson, L. K., Self-checking Digit Concepts, Journal of Systems Management, Vol. 25, no. 9, p 36-42 (Sept. 1974).

[7] Reitwiesner, G. W., Policing Transcription and Transposition Errors by a Single Check Digit Modulo 10 and 11, National Bureau of Standards Report, 10886, p 2 (June 30, 1972).

[8] Whiting, J. S., An Efficient Software Method for Implementing Polynomial Error Detection Codes, Computer Design, p 73-77 (March 1975).

[9] Mason, J. O., Connelly, W. E., The Application and Reliability of the Self-Checking Digit Technique, Management Advisor, p 27-34 (Sept. - Oct. 1971).

The Standards Implications of the Developing Interrelationships Between On-Line Bibliographic Retrieval, Data Manipulation and Micrographics Display

Robert M. Landau

Science Information Association
3514 Plyers Mill Road
Kensington, Maryland 20795

A brief history of the rapid development of the on-line scientific and technical information retrieval system (OLSATIRS) with emphasis on compatibility, convertibility and standards problems is described. A similar description is given for the developments in the areas of data and micrographics. Procedural, nomenclature, interchange and economic problems in these three fields are then considered. Comparisons, contrasts, and interrelationships between the three fields are provided. Problem areas and opportunities are suggested.

Key words: Standards; user interface; bibliographic information systems; numeric data systems; data; on-line systems; language; retrieval; information organization; micrographics; economics.

1. Introduction

The standards implications discussed in this paper are considered primarily from the viewpoint of the user and as a human/system interface problem in terms of language. There are many types (or levels) of language: source, input, indexing, classification, coding, machine, programming, search and retrieval, output and display.[1] This paper is concerned with only the last three and the related challenges and opportunities presented by the rapidly evolving changes in the storage media and retrieval techniques.

The needs and demands of the users of information systems have not changed. The information seeker still wants to be able to rapidly select and be provided (have displayed) as soon as possible desired information as required in an easily readable form at the lowest possible cost. It is well recognized that, although there is an avalanche of information being created, it is rare that an information seeker obtains more than a small fraction of the information he should receive from the system were it functioning at 100% efficiency and effectiveness.

Those who are systems oriented and perceive the advantages of rational information system organization and procedures point out that the closer we move along the spectrum from cooperation, convertibility and compatibility to standardization, the more effective and efficient will be our information systems. Most agree, but many point out that the transition must be paced so that innovation can be allowed to produce the optimum mechanisms, techniques and arrangements.

---

[1] Cooperation, Convertibility and Compatibility Among Information Systems, NBS Misc, Pub #276, June 15, 1966 pg 98.

2. Discussion

In the field of scientific information, the recent (5 years) proliferation of the information delivery media of on-line interactive retrieval systems (OLIRS) and micrographic systems (MS) have had significant impact on bibliographic information systems (BIS), but little impact on numeric data information systems (NDIS). A question frequently asked is: Has this been true because of(lack of) the imposition of standards in the respective fields and/or media? There is no single or simple answer to this question, but this paper considers some significant aspects of this question.

In 1970, there were only two major scientific and technical information (STI) data bases (sponsored by NLM and NASA) of a few hundred thousand bibliographic records each and a few small (a few thousand records each) STI data bases available from OLIRS to a few dozen people. In 1975, a preponderant proportion (80-90%) of all bibliographic STI produced since 1970 is available in one or more OLIR systems at costs ($10-$20 per search) significantly lower than computer batch or manual systems. A significant proportion (about 50%) of all original STI journal publication material (and over 90% of all STI report literature) is available in micrographic form at cost to the user of 1/5 to 1/2 of the hard copy material.

In contrast, the amount of S&T numeric data available in either OLIRS or micrographic form is trivial. Some questions relating to why this might be, include:

1. Are there inherent differences in form or content between BIS and NDIS to explain this contract?
2. Has emphasis on standards in one or the other field been an influencing factor?
3. Has consumer demand (or lack of it) caused the difference?
4. Have production and distribution costs been significantly different?
5. Are there contrasting user/system interface requirements?
6. Is the difference in the intellectual organization of material between the two fields important?

Although there is an abundance of published literature on BIS and on DMIS, there is little in the published literature to provide clues to the answers to these questions concerning comparisons between the two fields.

In my opinion, the answers to the above questions are:

1. Yes, because there is no comparable document surrogate (eg.bibliographic material including an abstract in BIS) in the NDIS field. Thus, the amount of information to be created and stored in machine sensible form in BIS is only a fraction of the full document (record) amount. In addition, the nature of data is such that it must be refined and confirmed over the years to the point that it is reliable and confirmable. This dynamic quality over time provides a difficult panorama of constantly shifting relationships between data.[1]
2. No, because there has been reasonably similar amounts of standards efforts in both fields.
3. No, the demand by users has seemed to be about the same in both fields.
4. Yes, the cost of producing and distributing bibliographic record (even including the abstract) in either printed hard copy (HC) or machine sensible (MS) form is much less than producing and distributing the full document (record), no matter what the form (HC or MS) or content (S&T reports and publications or numeric data).
5. Yes, the BIS on-line user interface requirements can now be

---

[1] Reference data on Thermophysics, Y. Touloukian, 4th International CODATA Proceedings, June 1974.

satisfied with relatively inexpensive alpha-numeric terminals (either CRT or printer) connected to one of several computer systems (through telephone dial-up time sharing networks) that now have well developed, powerful, boolean logic oriented software capabilities. However, a number of BIS search and retrieval languages have proliferated in recent years, a very disturbing trend which will be discussed in detail later in this paper. In contrast to the BIS, the NDIS user interface requirements include more expensive graphic terminals (either CRT or printer) to generate charts, graphs, tables, etc, and powerful graphic/mathematic/statistic oriented software in large time shared computer systems. Although there are many such software systems that are usually directed to a particular application, few are designed to handle large amounts of numeric data in time sharing interactive systems. However, the situation is now such that appropriate, but modest investments in the creation of on-line NDIS, could produce significant improvements in this field.

6. Yes, BIS are organized intellectually by classification schemes, thesauri, key terms, etc. and the user should understand not only the subject matter, but also the intellectual organization of the information and the resultant required search strategy and language. In contrast, NDIS data elements relate to each other in arrays, matrices, tabular form, mathematical relationships, time series, applications, sources, transforms and other heterogenous relationships. No significant body of methodological knowledge has developed concerning compatible record elements, intellectual interrelationships, on-line retrieval methodology or user/system interface language in the field of NDIS.

Aside from the above observations, it is felt that there are three other points to be made in contrasting the development of BIS and NDIS:

1. In the last ten years, the NSF has provided many millions of dollars for the creation of large S&T primary publishing activities and the associated secondary publishing activities.

2. As a consequence of this and natural technical progress, by 1970 or so, most primary and secondary publishers were producing their output in MS form on magnetic tape. Thus, the stage was set for entering the bibliographic material into on-line retrieval systems. But, in the early 70's many thought that since there were no standards for bibliographic records (and, indeed, no two organizations output were alike!) it would be very difficult, time consuming and expensive to convert the different formats into a common format for on-line storage and retrieval. The fears were groundless; this was a trivial programming problem. Therefore, by 1975, essentially all current and recent S&T information is now available on-line in a convenient and cost effective manner but through a number of different query languages from a number of different computer systems.

3. Since 1972, there has been aggressive commercial marketing of the on-line S&T BIS and constant attention paid to user needs and requirements.

None of the three above points can be made concerning the NDIS. There has been a comparatively modest amount of NSF (and other) funding for NDIS; there is very little numeric data in MS form and, consequently, very little commercial marketing of on-line NDIS.

An example of the complexities of the "standards" problem in this field of human/systems interface can be provided by describing a short study performed by a group with which I am associated. A comparative analysis of methods to simplify the human/machine interface for on-line interactive retrieval for S&T BIS was conducted.

The rapid proliferation of a bewildering array of commands required by the various systems was noted. See Chart 1 for examples of only five of more than 30 known systems all having differing commands for identical functions. Numerous experts have stated that a full context natural language is the best and most desirable human/system interface language. All system designs are moving toward this "highest" level language - in the U. S. - English.

With the above thoughts in mind, the group considered several possible scenarios:

A. Continuation of the present state of affairs - the proliferation of more and more discreet command languages.

B. Work toward agreement of the use of identical commands for identical functions.

C. Create and provide an interface language link program inside and between each computer/communication system.

D. Create and provide a full natural language (English)interface to be interposed between every user and each on-line retrieval system.

Each of the above scenarios were considered in terms of the following factors:

1. System considerations
2. Human factors
3. Timing required for creation and acceptance of change.
4. Economic considerations

Scenario A is the easiest from the viewpoint of individual system development. No coordination (and consequent system modification) would be imposed from sources external to each of the developing system. From the user viewpoint, when two or more command languages must be learned, a saturation tolerance point is soon reached. Assuming a potential one million users of on-line S&T BIS, it was estimated that less than one half would ever routinely use even one formatted command language. The other half would either use an information expert intermediary or not use the system at all. It was further estimated that less than one half of the user half would tolerate two formatted command languages; less than 1/2 of that 1/4 will tolerate three command languages, etc. Now the factors may be 1/4 or 3/4, but the point is that under this scenario, very few end users would be using the systems. Most access would be through intermediaries. The group concluded that there might be a slow consensus of common commands agreed to by the system operators over the years,but actions to date provide little optimism for this possibility. The overall economics of this scenario were deemed to be clearly unfavorable. The unnecessary training costs for hundreds of thousands of users would be staggering and it was concluded that most "trainees" would be "dropouts" (eg. non-users) after little or no actual system use. The amount of frustration and loss of efficiency would be (probably already is) hard to measure, but certainly is a significant factor.

Scenario B could be considered a sub set of A and, in effect, is being tried at various levels right now. However, based on past experience, only modest success at best, was visualized by the group. When (if ever) we all agree to LOGON-LOGOFF or LOGIN-LOGOUT or whatever, I will be pleasantly encouraged to expect even greater "cooperation."

Scenario C had considerable attraction at first blush. It is relatively simple to accomplish from a system sense; it would require that each user be trained in only one command language, but yet use all other systems; it would be relatively simple to implement and be accepted by the users and would not be unreasonable expensive either in initialization or continuing operation.

But all these statements were proposed with the assumption that there would be economic incentives to the operating time sharing networks and that there would only be two or three of them. Neither assumption proved to be true. Networks are proliferating broadly and rapidly and even assuming that the operators were willing to implement such software packages, the operational overhead would rise geometrically with the number of networks; eg., 90 interlink programs would be required for 10 networks. A little further calculation easily persuaded us that this approach 1) was not economically feasible 2) was not likely to be implemented even if it were and 3) still required all users to be trained in one formatted command language.

Scenario D appeared upon early consideration to be a difficult technical problem; artificial intelligence was not that developed; several major attempts had been made to develop free English input systems with no successful commercial development resulting therefrom. The technique appeared still to be in the experimental stages. From the users viewpoint, however, it was agreed that this is the best solution. Little or no training would be required; the user base could broaden to its natural maximum. High system use would drive down unit search costs; the change would be minimal and easily accepted. No intermediaries would be required. If such a package could be developed, implementation would be relatively easy with minimal system perturbation. After a careful review of this field, it was determined that artificial intelligence had, indeed, progressed sufficiently to provide a high probability fo successfully producing a software package to provide a practical and useful full natural language interface that could be interposed between the on-line user and the Data Base Manager (DBM) of any typical medium to large scale computer of the type now being used by all the major data base vendors.

Based upon the information briefly summarized above, the study group concluded that, on balance, scenario D was the most attractive long term option considering the four factors described above.

CHART 1

ON-LINE COMMANDS

A Quick Users Guide for Bibliographic Search Systems

Compiled by

Barbara Lawrence (Exxon Research and Engineering Co., Linden, N.J.) and
Barbara G. Prewitt (Rohm and Haas Co., Spring House, Pa.)

May, 1975

FUNCTIONS	ORBIT	ELHILL	DIALOG	RECON	TYMFACT
HOUSEKEEPING					
Starting search after login	Directly connected	Directly connected	BEGIN	BEGIN	Directly connected
	N A	"RESTART"	BEGIN (insert file number) ●FILE (insert file number)	N A	BEGIN
Terminating a search and disconnecting from system	"STOP"	"STOP"	END LOGOFF	BYE	STOP LOGOFF QUIT
Asking for list of accessible data bases	"FILES?"	"FILES?"	? FILES BEGIN	BEGIN	BEGIN
Determining elapsed time	"TIME" "TIME INTERVAL" "TIME RESET"	@ (Time given at login and logoff)	N A — included in END command	END	TIME
Deleting search statements no longer needed	"ERASEBACK" "BACKUP" "RESTACK" "ERASEALL"	"ERASEBACK" "BACKUP" "RESTACK" "ERASEALL"	N A	RELEASE	N A
SEARCHING					
Entering search terms	Enter words (always in search mode; can search on single terms or multiple term concept)	Enter words (always in search mode; can search on single terms or multiple term concept)	SELECT (Simple form can search one word, multiple term concept or term(s) from EXPAND display)	Enter words (always in search mode; can search on single terms or multiple term concept)	SELECT (Simple form can search one word, multiple term concept or terms(s) from EXPAND display)
	"FIND (used for by passing program's queries)	"FIND (used for by passing program's queries)	N A	N A	N A
Displaying an alphabetical list of terms	"NEIGHBOR (displays up to 10 terms; ability to go forward or backward thru index)	"NEIGHBOR (displays up to 10 terms; ability to go forward or backward thru index) "NEIGHBORDET (displays headings and needed sub-headings 5 at a time)	EXPAND (displays 20 lines; shows existence of related terms in a thesaurus)	EXPAND (displays 10 terms; user can modify)	EXPAND (displays 10 terms; user can modify)
Creating search logic	Always in search mode. Insert AND, OR and AND NOT in any search statement	Always in search mode. Insert AND, OR and AND NOT in any search statement	COMBINE must use set numbers with AND, OR, and AND NOT logical operators	Always in search mode. Insert AND, OR, and AND NOT in any search statement	COMBINE must use set numbers with AND, OR, and AND NOT logical operators
Changing data bases	"FILE (insert file name)	"FILE (insert file name)	● FILE (insert file number)	BEGIN (insert file number)	BEGIN (insert file number)
Root Searching	**TERM #** (for single character) **TERM** (for multiple characters)	**TERM #** (for single character) **TERM** (for multiple characters)	SELECT **TERM?** (up to 50 terms)	**TERM**	SELECT **TERM**

N A Not Applicable

CHART 1 - contd.

FUNCTIONS	ORBIT	ELHILL	DIALOG	RECON	TYMFACT
SEARCHING CON'T					
Text searching	STRINGSEARCH (searches for words or imbedded character strings) SENSEARCH (searches for words or imbedded character strings in a sentence)	STRINGSEARCH (searches for words or imbedded character strings) SENSEARCH (searches for words or imbedded character strings in a sentence)	SELECT **TERM**(w) **TERM** (word proximity search, different forms available)	**TERM TERM** (different forms available)	SELECT **TERM**(w) **TERM** (word proximity search, different forms available)
Restricting searches	Date ranging	Date ranging	LIMIT LIMIT ALL (dates, language, accession numbers, special features)	LIMIT LIMIT ALL (dates, language, accession numbers, special features) **TERM A TERM B** (range searching)	LIMIT LIMIT ALL (dates, language, accession numbers, special features) SELECT **TERM A TERM B** (range searching)
OUTPUT COMMANDS					
On-line printing	"PRINT" "PRINT TRIAL" "PRINT FULL"	"PRINT" "PRINT DETAILED" "PRINT FULL"	DISPLAY (primarily for CRT) TYPE (primarily for hard copy terminals)	DISPLAY (primarily for CRT) TYPE (primarily for hard copy terminals)	DISPLAY (primarily for CRT) TYPE (primarily for hard copy terminals)
Formatting printout	"PRINT ____, ____ ____, ____" (specify search statement number(s), format instructions and field acronyms; prints as directed)	"PRINT ____, ____ ____, ____" (specify search statement number(s), format instructions and field acronyms; prints as directed)	N.A.	FORMAT (specify field acronyms to define format; output command with reference to Format 4 must be entered subsequently)	FORMAT (specify field acronyms to define format; print command)
Off-line printing	"PRINT OFF-LINE" (also applies to any PRINT command above)	"PRINT OFF-LINE" (also applies to any PRINT command above)	PRINT	PRINT	PRINT
Sorting	N.A.	N.A.	SORT (applicable only to selected data bases and fields)	● SORT	SORT
Interrupting on-line output	N.A.	N.A.	break key	break key	break key
Erasing single characters	backward slash key or c/ (varies with terminal type)	backward slash key or c/ (varies with terminal type)	back arrow or backspace key (varies with terminal type)	control key and H	back arrow key
Erasing whole lines	dollar sign key	dollar sign key	escape key	break key	escape key
SUPPORT FEATURES					
Requesting system news	"NEWS"	"NEWS"	"NEWS	N.A.	N.A.
Providing explanation of commands, program messages, operating procedures, etc.	"EXPLAIN ____" "EXPLAIN" (refers to last command)	"EXPLAIN ____" "EXPLAIN" (refers to last command)	EXPLAIN ____	HELP ____	HELP ____
Assisting user on how to proceed	"HELP"	"HELP"	N.A.	HELP	HELP
Personalized assistance	"COMMENT"	"COMMENT"	SEND MESSAGE/	Hot line assistance provided free of charge	N.A.
Providing description of search	"DIAGRAM" (search logic)	"DIAGRAM" (search logic)	DISPLAY SET HISTORY	HISTORY	HISTORY

N.A. = Not Applicable

Additional copies available from:
The National Federation of Abstracting and Indexing Services
3401 Market Street
Philadelphia, Pennsylvania 19104
Prepaid: $1.00

Product Coding-One Number from Maker to User

John T. Langan

Director of Systems Development
Distribution Codes, Inc.
401 Wythe Street
Alexandria, Virginia 22314

INTRODUCTION

The United States economy has passed from the industrial age into the age of distribution. Each year the physical volume and dollar value of the flow of commodities to market increases. Commercial channels of distribution are straining to place more than $800 billion of goods, produced by some 400,000 manufacturers, at the precise point of user need at the time needed.

That tremendous flow of goods is comprised of tens of millions of different products. A standard numbering system used for unique product identification can help accomplish this enormous task with increased efficiency in the operation and the functions of a company.

One of the goals for any company in standardizing product identification systems is to get the maximum use of data resources and to create a more effective means of collecting and exchanging product information with others. Standards for unique product identification are consensus agreements between the sender and the receiver. Before meaningful exchange of data can take place, there must be understanding and agreement on the method of identification, what it means, and how the product information will be represented.

THE NEED FOR A NUMBERING SYSTEM

Today, there is a broad, sophisticated and continuously proliferating product offering. Many manufacturers are making basically the same product, but with subtle differences. In order to give products specific identification without endless repetition of long and cumbersome descriptions, the age of distribution requires a standand unique product identification system. It requires a language that people can use to communicate with a machine, that a machine can use to communicate with people, and that people can use to communicate with each other.

There are many different numbering systems designed to do special jobs. Manufacturers use alpha and/or numeric systems to indicate how the product is fabricated, what it is composed of, and where the components are located within a plant. Some of those numbering systems, designed to assist production rather than distribution, have 35 or more alphanumeric characters. Numbers identify production runs, food packs, shelf life, perishability, etc. Social Security numbers identify people; bank account numbers identify depositors; and credit card numbers identify holders. The list of incompatible numbering systems continues to grow.

The distribution function requires only one number to get a product from point of production to point of consumption or use. The smaller that number, the more economical the system as a whole. A relatively simple identification of a product's "name and address" will lead to an efficient system designed to meet the total needs of the distribution process.

WHY AN ALL-NUMERIC CODE?

Accuracy is the reason for an all-numeric code. Telephone companies have spent millions of dollars converting telephone numbers from alphanumeric to all-numeric.

People read numbers more accurately than letters and/or combinations of letters and numbers. In typing or keypunching, there is less chance for error if only combinations of 10 numbers are used as compared to 26 keys needed for letters or 36 for numbers and letters. Within those systems utilizing forms which carry product identification information in the form of catalog numbers, etc., difficulties arise when mixing numbers and letters together. This problem is heightened in those cases where information is entered manually and the receiver endeavors to distinguish between an 8 versus a B, a zero versus the letter O, a Z versus a 2, etc.

THE DC SYSTEM

The system developed for uniquely identifying products throughout the distribution process is the Distribution Code (DC). Distribution Codes, Inc. (DCI) administers the Distribution Code for the distribution industry. The DC product numbering system provides an accurate, efficient and economical way of controlling the flow of goods throughout the entire commerical distribution process.

Sponsored by the National Association of Wholesaler-Distributors (NAW), representing some 30,000 individual wholesaler trade companies, the Distribution Code is supported by the following NAW member national associations:

- National Association of Electrical Disbributors
- Air-conditioning & Refrigeration Wholesalers
- Automotive Service Industry Association
- National Electronic Distributors Association
- National Industrial Distributors Association
- Southern Industrial Distributors Association
- National Welding Supply Association
- Northamerican Heating & Airconditioning Wholesalers
- Wholesale Stationers' Association

These associations jointly represent industries with some $81.3 billion in annual sales at the merchant wholesaler level.

Early versions of the DC came into widespread use in 1965 by distributors of electrical equipment and supplies and medical and surgical devices and supplies.

The DC is an 11-digit, all-numeric code, using 6 digits to identify the manufacturer and 5 to identify the manufacturer's item. It has ample capacity for the 400,000 manufacturers which research showed might need to be accommodated. DCI is responsible for complete control of all manufacturer identification numbers to prevent duplication and provide correct identification. The 5-digit product code of the DC will accommodate 100,000 items produced by each manufacturer. If a firm produces more than 100,000 products, an additional manufacturer number can be assigned.

While the DC is in widespread use in the non-retail sectors of the distribution function, it is by design that the Universal Product Code (UPC), used with consumer products sold through grocery, drug and mass merchandise/discount retail outlets, is totally compatible with the Distribution Code. For example, the use of DC manufacturer identification numbers starting with zero is strictly controlled, so as to avoid any duplication with UPC numbers which may also start with zero and are also 11 numeric digits long.

DCI assigns the unique "manufacturer" part of the DC identification number. The task of compiling and/or assigning the "item number" portion of the DC is a tremendous one which is being done by a number of third parties. These include agents in the following industries:

- o Electrical
- o Heating, Airconditioning & Refrigeration
- o Industrial Supplies and Equipment
- o Plumbing

These agents make the DC number available through directories, catalogs, various electronic data processing (EDP) media, pricing services, etc. for established commodity lines.

The sale of such services generates the revenues necessary to establish and maintain the DC System in perpetuity. This places the modest cost of maintaining the system on those who benefit most from its existence.

ADVANTAGES OF THE DC SYSTEM

One example of increased efficiency through the DC System is in ordering or reordering products. With other methods, an order clerk may write:

144 60 watt, white, frosted, standard-base light bulbs
The XYZ Electric Mfg. Co.
Lamp Division
7516 Goshen Street
San Francisco, CA 90101

With the DC System, the clerk could have reordered all that data far more simply by writing:

144 123456-12345

The 11-digit DC number becomes a complete and unique designation for one, and only one item.

Neither the identity of the manufacturer nor the identity of his products are impaired by the DC System. The manufacturer's name and logo still appear on boxes, cartons, cases, display packs, items and so on. Manufacturers' catalog numbers, some of many years' standing and widely recognized in the industry, are still preserved in his catalogs. However, many manufacturers also show the 5-digit DC item number in their catalogs.

ADVANTAGES TO THE MANUFACTURER

The DC System is an easy, low-cost method of product numbering. Sequential numbering is simple to construct and seldom needs revision. With products identified by relatively short and widely used numbers, a manufacturer will not be plagued with requests from customers to place "their number" on the product.

Today, in a number of industries, most orders go to a manufacturer with only the DC number included. A simple computer conversion will relate the DC number to a manufacturer's catalog or production code which may have as many characters (alpha or numeric) as a manufacturer needs. Some manufacturers use the DC number for internal purposes as well, thus eliminating conversion costs.

The DC System permits orders to be communicated directly from a whole-saler-distributor's customer to the distributor and to the manufacturer's computer, thus reducing paperwork and chances for error and greatly reducing response time throughout the distribution process.

ADVANTAGES TO THE WHOLESALER-DISTRIBUTOR

When a wholesaler-distributor begins to use a computer, he realizes that each item in his inventory must be assigned a unique number. Assigning and maintaining these numbers is an essential, time-consuming and expensive initial set-up task.

However, if a wholesaler-distributor has been using DC numbers in his manual system to save time and errors in purchasing, the need to assign numbers is eliminated. His employees are already accustomed to using the DC number and do not need to be retrained. This means fewer headaches during the conversion process and faster realization of the benefits of computerization.

When the DC number is printed on the item (where possible), package, carton and case, then order taking, receiving, picking and shipping will all be greatly simplified. In some cases (such as hardware, for example) the productive selling time of outside salesmen can be significantly increased. Instead of spending time consulting voluminous descriptive and illustrated catalogs to determine the catalog number, the salesman can write orders directly from the DC number printed on the items in the customer's stock or "want" book.

ADVANTAGES TO RETAILER/DEALER/BUSINESS USERS

With the DC System coming into greater general use, there is a strong tendency for retailers, dealers and business users to utilize those numbers for product identification. This reduces the cost of paperwork associated with purchasing, and increases accuracy.

Instead of writing a lengthy description of the item needed, they simply write the 11-digit DC number in their "want" books.

COST-EFFECTIVENESS

The DC System provides for increased accuracy, decreased paperwork and expeditious flow of supplies. In turn, this offers the opportunity of maintaining smaller inventories within the channels of distribution. The DC System thus provides the opportunity for increased economies throughout the entire distribution function. Performing the distribution function better, faster and at a lower cost ultimately benefits the consumer, the end user who must bear the total cost of distribution.

Development of a Data Dictionary/Directory Using a Data Base Management System[1]

E. K. C. Lee and E. Y. S. Lee[2]

Jet Propulsion Laboratory
Pasadena, California 91103

The Data Dictionary/Directory (DD/D) was initially developed as the nucleus for a large data system called the Military Construction Data System (MCDS) at the U.S. Army Construction Research Laboratory (CERL). The DD/D allows nonprogrammers, such as engineers and scientists, as well as analysts easy access to information on the characteristics of the data systems. Using a general functional approach to the design of the DD/D, important functional and support requirements of the DD/D were extracted and then factored into the structure of the DD/D as five different facilities. Using a Data Base Management System (DBMS) called System 2000[3], the complete structure was directly translated into a hierarchical data base structure using the data base definition facility of the DBMS. Convenient and flexible retrievals in the form of reports and matrices can be readily obtained using the predefined retrieval commands called STRINGS.

A second implementation of this DD/D on an IBM 370/158 with VS2-R1.6 is being developed at the Jet Propulsion Laboratory (JPL). Because of the flexibility of the initial design, only minor changes in data structure and retrieval commands are needed. The data systems involved at JPL are completely different from the initial implementation at CERL, with data bases in areas such as finance and accounting, business, and personnel applications.

Key words: Data base management; data base management system; data definition; data dictionary; data directory; data element; data structure, data system; Federal standard; information system; standardization; System 2000.

---

[1] The initial research and development of the DD/D was performed while the authors were with the U. S. Army Construction Engineering Research Laboratory, Champaign, Illinois, 61820. The modification of the DD/D for JPL use was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract NAS7-100, sponsored by the National Aeronautics and Space Administration.

[2] ADP Analyst and Member of the Technical Staff, respectively.

[3] Product of MRI Systems Corp., Austin, Texas, 78766.

1. Introduction

In early 1973 a decision was made to consolidate all the construction-related data bases at the U.S. Army Construction Research Laboratory (CERL) into one single data system called the Military Construction Data Systems (MCDS) [1][4]. The data bases involved were in various stages of development, some had been in operation for several years and some were being implemented using the DBMS System 2000[5]. During the implementation of MCDS it was discovered that not only the important roles of the Data Base Administrator (DBA) needed to be carefully defined [2], but a new tool called Data Dictionary/Directory was required to help the DBA to effectively perform his/her duties. The importance of a DD/D and its various functions have been discussed by many including description of some development efforts [3], [4], [5]. Meanwhile, developers of Data Base Management Systems (DBMSs) have begun to realize that some of the functions of the DD/D can be incorporated into the DBMS and some DBMSs do provide supports such as a directory and data element definitions. We feel that an independent DD/D has several advantages over the one integrated with a DBMS. For example, an independent DD/D managed by a DBMS can handle many diversed functions much more comprehensively, and at the same time it can form the nucleus of a large data system consisting of many data bases that may or may not be managed by the same DBMS.

In the course of the development of the MCDS and its DD/D we found many recurring problems that can be effectively solved by providing either the facilities or supports in a DD/D. Some of these problems were:

(1) Data redundancy.

(2) Lack of standardization.

(3) Lack of sources or derived information of the data.

(4) Lack of comprehensive description of data.

(5) Lack of tools to make estimates on the effort of changing the data base.

(6) Lack of centralized tools for the DBA to effectively carry out his/her duties.

The above were only some of the major problems we encountered during the implementation of the MCDS, but they formed the basis for the development of a DD/D that would solve all of the above problems and provide additional benefits. The second implementation of the DD/D with minor improvements is being carried out at the Jet Propulsion Laboratory (JPL). Here we are concerned with the development and conversion of the data bases in the Administrative Computer Section (ACS) with emphasis on finance and accounting, personnel, and business applications. We shall discuss the general design approaches and the structures and facilities of the DD/D in the next two sections.

2. Design Approaches

---

[4] Figures in brackets indicate the literature references at the end of this paper.

[5] Product of MRI Systems, Corp., Austin, Texas, 78766.

2.1. Functional Design

After making a survey of the various available DD/D [3] as well as our own requirements, we came up with a list of over 20 various functions or supports of a good DD/D. We condensed this list into four major categories: (1) record keeping, (2) cross-referencing, (3) indexing, and (4) standardization and control. We will discuss them separately:

(1) Record Keeping. Any good DD/D will provide a means of keeping an order list of all data elements, their characteristics, attributes, synonyms (if included), and data definitions. These records should provide: (a) a vehicle to identify common data items among the various data bases in the data system, and (b) a complete machine readable and processable representation and documentation of all data elements.

(2) Cross-Reference. Besides record keeping, the DD/D should provide the following six categories of cross-reference between the data elements and input/output records. They are the cross-reference between data elements and: (a) various synonyms, (b) users or related systems, (c) files (input and output), (d) data sources, (e) reports, and (f) programs and special routines. Using these cross-reference tools, the programmer and analyst can determine the data relationships and data paths for both storage and retrieval. These cross-references are valuable design tools to minimize cost for designing new reports, new systems, or changes in old ones, and can also provide more reliable estimates of cost and manpower for these tasks.

(3) Indexing. To allow both nondata processing personnel as well as data processing personnel to use the DD/D effectively, a good indexing scheme must be provided. There are two general approaches in indexing schemes, i.e., to relate the data elements to their definitions: (a) key word in context (KWIC), and (b) standardized indexing terms. Each method has its advantages and disadvantages and the implementor of the DD/D should make the judgment according to the specific needs of the installation. However, the standardized key word scheme does provide a means to develop a standard that can be gradually expanded to other parts of the DD/D. We chose the standardized key word scheme also as a first step to implement the requirement of the Federal Regulations [6]. The standard key word list for MCDS was taken from 'Thesaurus of Engineering and Scientific Terms', U. S. Department of Defense, 1967. A new version of this document has just been published [7].

(4) Standardization and Control. The implementation of the DD/D will allow the DBA to exercise strict control and impose standards over all the data elements in the data systems. For example, the DBA can institute a procedure whereby all new data elements must conform to a standard governing the definition, validation, etc., of the data elements before they can be entered into the data system or any data bases. Similarly, the DBA can limit the access of certain data elements by enforcing certain security procedures. The DBA may also exercise control over the collection of source data, creation of new data, multiple use of old data, capture of data from other sources, and even the design of new data bases from other sources, and associated applications within the data system. At a Federal Agency, this is also the first step toward fulfilling the Federal Regulations in the standardization of data elements and representation [7].

2.2. Use of DBMS

The above four major functions represent many forms of data and textual materials, and from the start we knew that we needed a DBMS to handle these diverse forms of information. The structure of these functions lends itself readily to a hierarchical form that can be easily described using the data base definition facility of System 2000. The idea of using a DBMS to manage a DD/D is not unique with our development; similar approaches have been taken by Eastern Airlines in developing the EAL Data Base Directory System using TOTAL [3] [8]. A more specific application using System 2000 is being utilized by the Texas Governors Office of Information Services [9]. Our DD/D can probably be implemented by any DBMSs that support a hierarchical data base structure.

Here are some of the advantages of using a DBMS to maintain a DD/D:

(1) Structure of DD/D can be very flexible and can be changed readily for a given data system.

(2) Updates and retrievals can be readily performed using the facilities of the DBMS and by nonprogrammers.

(3) In the case of System 2000, over 30 predefined retrieval commands have been developed to meet the diverse needs of many types of inquiries from a complete dictionary request to a specific request on a project.

3. Structures and Facilities of the DD/D

The hierarchical structure of the DD/D can be divided into five levels that can, in turn, be divided into five facilities reflecting the four major functions and supports of a DD/D mentioned in the previous section. Figure 1 shows the data base structure of the ACS DD/D and the five facilities that are described separately as follows:

(1) Data Information Facility (DIF). This is the central point of the structure of the DD/D. Each entry in this facility represents a unique data element of the data system and contains all the information about the given data item such as data name, date of update, data abbreviated name, data attributes, and data definition. A complete data dictionary can be produced by listing every entry in this facility. Figure 2 shows the listing of the data dictionary. In addition, this facility is linked directly with the following three facilities: (a) Cross-Reference System Facility, (b) Data Code Facility, and (c) Key Information Facility. The DIF fulfills mainly the record keeping function of the DD/D.

(2) Cross-Reference System Facility (CRSF). In the ACS DD/D, this facility provides cross-reference matrices to four different input/output areas: (a) file, (b) report, (c) program, and (d) input source. Cross-references are also provided for data synonyms, data codes, and common data elements among different systems. This information is of particular value, since it shows the user how the data from different systems are basically the same. See figure 3 for a report on the cross-reference data elements for all systems. Using the available STRING commands, a wide range of reports relating to the CRSF can be obtained for each user application or system. In combination with the SPAN function of the STRING command, only a selected range of data items will be retrieved without examining the cross-reference for every data element. Figure 4 shows an example of the cross-reference of a

report matrix for the Travel Advance Account System (TAAS). Besides fulfilling the cross-reference function, CRSF also provides support for the control and standardization function.

(3) Data Code Facility (DCF). Linked with the Data Information Facility is the Data Code Facility, which is an important tool for the DBA and used to enforce a standard code that is applied to all unique data elements. For example, each Army installation is usually assigned a unique code to distinguish one installation from another. Due to various applications over the years, several sets of codes have been in use. The DBA can select or create a set of standard codes to be used by all users of the data system and this set of codes will be stored in this facility. DCF supports both the major function of record keeping as well as control and standardization.

(4) Key Information Code Facility (KICF). As mentioned before, we chose to use a standard list of key words to index each data definition. To simplify the indexing scheme, we assign a key information code to each key word (or key phrase), so that one or more key codes can be linked to each data element and all the associated information in the Data Information Facility. The actual key word list is stored in the Directory Facility. Besides supporting the major DD/D function of indexing, KICF also provides support for the control and standardization function.

(5) Directory Facility (DF). Since each of the above facilities uses extensive codes to simplify the representation and storage of such information as file, report, program, and input source, a directory is needed to link these codes to the actual name or descriptions. Whenever possible, the actual names on the directory are used to produce the necessary DD/D reports instead of the codes. The DF supports the major function of record keeping as well as the control and standardization function.

4. Implementation Using System 2000

The above data base structure of the DD/D can be readily achieved in the System 2000 DBMS using the defining language of the Define Module of of System 2000. Each of the boxes in figure 1 is represented by a Repeating Group (RG) in the data base definitions. The zero level RG is the system information that is the root of all subsequent RGs for a given system, such as the TAAS mentioned before. The first level RG is divided into two groups. The first group is a very large RG with one entry for each data element supporting the DIF. The other eight RGs belong to the Directory Facility. Three RGs are represented at the second level; all are linked to the DIF RG. Each RG represents the CRSF, DCF, and KICF respectively. The CRSF RG extends two more levels down. The last level is a way to handle overflow data at the previous level (level three), which consists of six RGs, each of which represents one of the six areas of cross-reference available.

The initial loading of data into the DD/D can be accomplished using the Load Module of System 2000 with the data in the "loader string formats." Subsequent changes and updates can be performed by either the Immediate Access or Queue Processing Module, both modules support a simple English-like language. Ad hoc retrieval can be performed, both on-line or via remote batch operations using the above two modules. However, a more powerful tool is the predefined STRINGS, which can be created using the Define Module and stored in the DD/D data base. Using these STRINGS, non-data processing personnel can obtain different reports, perform updates, and retrieve diverse information by invoking the STRING with substitution of specific parameters during execution.

5. Some Implementation Experiences

Because of the special requirements at each installation, the implementation of the DD/D has been customized. Certain functions and supports of the DD/D are emphasized during the development cycle of the data systems. The general methodologies used in applying DD/Ds to the development of an information system were discussed by Sibley and Sayari [10]. We will discuss the implementation of the DD/D as it is being utilized in the development of MCDS at CERL, and the development of business and administrative data bases at JPL.

5.1. Implementation at CERL

The MCDS DD/D is being implemented at CERL using CDC6400 Kronos computers in the Control Data Corporation (CDC) Kronos Time Sharing Computer Network. Both on-line and remote batch operations are supported. The objective of the MCDS is to develop and maintain a long-range outlook of data system requirements for planning, engineering and designing, construction, operation, and maintenance of Army-constructed military facilities [1]. The major function emphasized in implementing the MCDS DD/D is, therefore, to standardize common element names, types, definitions, codes and input, and access procedures. The DD/D is also the information tool for engineers, researchers, and analysts to find out if the data they need are available in some data bases within MCDS before they proceed to expend money and manpower in collecting data from the field or performing some complex computation to arrive at derived data.

Several new data bases are being developed using the DD/D at CERL; two of them using the System 2000 DBMS are the Environmental Impact Computer System [11] and the Life Expectancy of Facilities [12]. To aid in the development of these new data bases, the DD/D is: (a) the management tool coordinating the development of these new data bases, and (b) the design tool for the analysts to obtain an overview of the available data elements, files, reports, etc.

5.2. Implementation at JPL

The objective in implementing the ACS DD/D at JPL is somewhat different than that of MCDS DD/D at CERL. Although the fact that the DD/D is being used as a tool for managing, designing, and documenting the development of new data bases is still valid at JPL, additional objectives are: (a) to provide standardization and control of all data elements used in the finance, accounting, personnel, business, and administrative data bases and files, and (b) to convert the current applications into data bases operating in a DBMS environment using System 2000; centralizing some of the data sources and files into fewer data bases is a secondary objective.

The implementation at JPL uses an IBM 370/158 operating under VS2-R1.6. Because the DBMS System 2000 is being used in both implementations, the DD/D structure is readily transferable from one computer to the other with only minor changes.

6. Conculsion

A Data Dictionary/Directory has been developed using a Data Base Management System called System 2000. The advantages of using a DBMS to manage a DD/D are many; flexibility of the data structure and ease of updates and retrievals by nonprogrammers are the major ones. Using the 'STRING' facility of System 2000, over 30 retrieval and 20 update commands have been developed for convenient access and management of the DD/D. Because the DD/D is a data base itself (using the DBMS), a copy of it can be saved weekly and the updates are saved after every update transaction. The implementation of the DD/D at CERL is concerned with a large Military Construction Data System with many data bases in both an engineering application and a research environment, while at JPL the implementation is for data bases in the finance, accounting, personnel, business, and administrative areas. Regardless of the installation environment, the DD/D provides the following key benefits:

(1) Standardization of all data elements, data definitions, and codes.

(2) During the development phases of new data bases or conversion of old ones, the DD/D can provide:

  (a) Tools to manage and coordinate the development cycle.

  (b) Assistance to the analyst and DBA in the design.

  (c) Tools for documentation of the data system.

(3) The DD/D can be used as an information system to:

  (a) Avoid any duplication in efforts to collect field data or compute derived data.

  (b) Assist the conversion from common file applications into DBMS environment.

  (c) Study the feasibility of centralizing or decentralizing data bases in a data system.

7. References

[1] Lee, E. K. C., Development of the Military Construct Data System (MCDS), Part 1, CERL, NTIS Doc. AD/A-00710/4GI (Sept. 1974). (Available from National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22151).

[2] EDP Analyzer, The 'Data Administrator' function, EDP Analyzer, 10, 11 (1972).

[3] EDP Analyzer, The Data Dictionary/Directory function, EDP Analyzer, 12, 11 (1974).

[4] Granberg, F. and Hauger D., Anderson Clayton Foods Data Dictionary/Directory, Proceeding of the 37th Meeting of GUIDE International, pp. 161-177 (Nov. 1973).

[7] United States Documentation Center, DDCH 4185.7, DDC Retrieval and Indexing Terminology, AD-A001201 (1975).

[8] Collard, A. F., a data dictionary directory, J of System Management, 25, pp. 22-25 (June 1974).

[9] Eberle, S. L., England, L. D., and Schiff, B. H., The use of data base management system for standards analysis, edited Data Elements in Information Processing, COM74-10700, pp. 67-77 (Apr. 1974).

[10] Sibley, E. H. and Sayari, H. H., Data Elements Dictionaries for Information System Interface, ibidem, pp. 285-300, (Apr. 1974).

[5] Setzer, W. K., and Keen, J. M., The use of a data dictionary for IMS change control, Proceeding of the 37th Meeting of GUIDE International, pp. 161-177, (Nov. 1973).

[6] National Bureau of Standards, Standardization of Data Elements and Representation, Federal Information Pro-Standards Publication 28 (Dec. 1973).

[11] Lee, E. Y. S., Jain, R. K., Lee, E. K. C., and Goetel, B. Environmental Impact Computer System, CERL, NTIS Doc. AD-787295/5G1 (Sept. 1974).

[12] Lee, E. K. C., Kirby, J. G., and Grgas, J. M., An information Storage and Retrieval System for Life Expectancy of Facilities, CERL, NTIS Doc. AD-782912/OGI (Apr. 1974).

FIGURE CAPTIONS

Figure 1. Data base structure of ACS Data Dictionary/Directory System.

Figure 2. Output example of ACS Data Dictionary.

Figure 3. Output example of cross-references of common data elements in systems (projects).

Figure 4. Output example of cross-reference report matrix for all data elements in a given system (part of output of TAAS is shown).

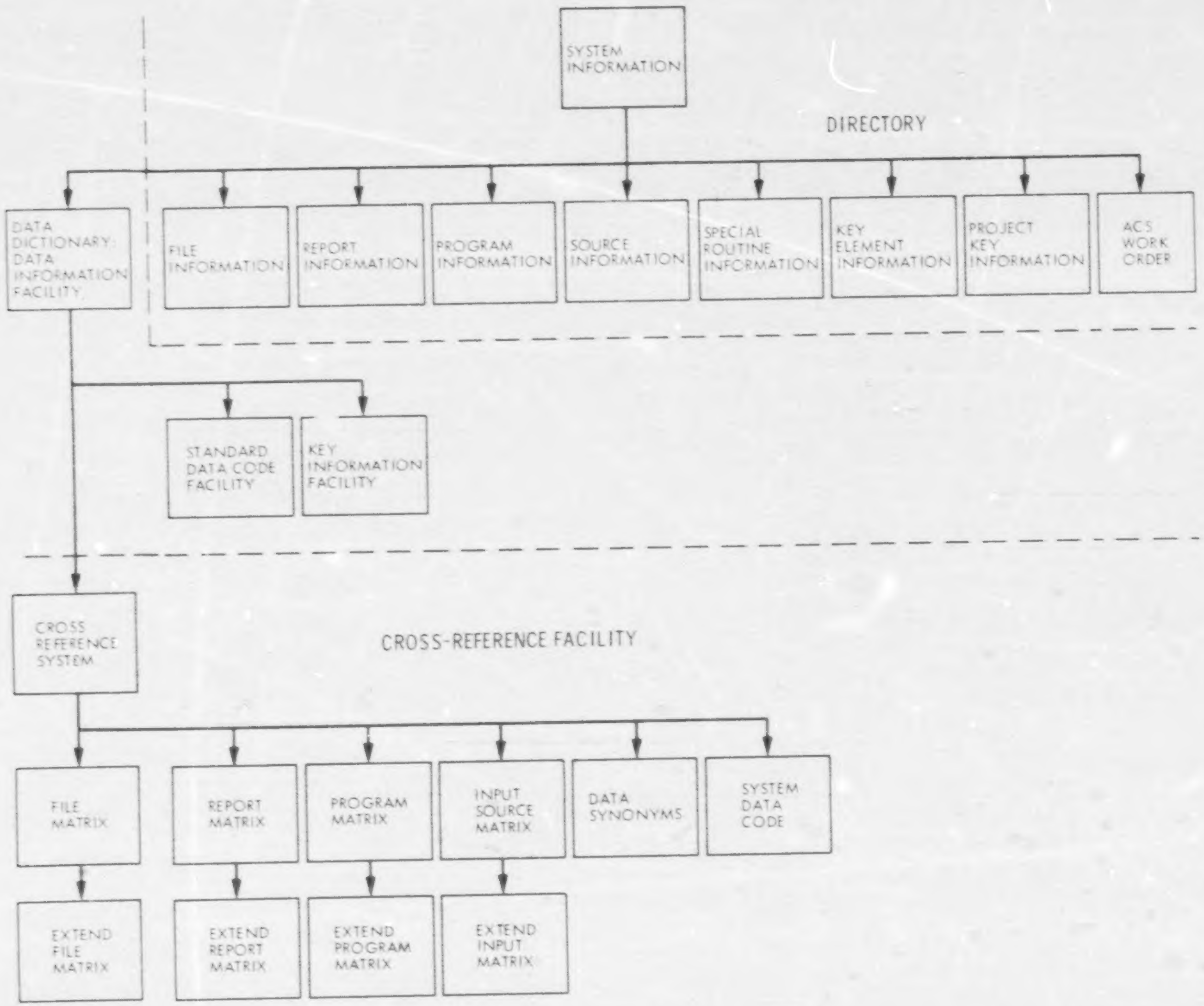

SYSTEM INFORMATION
DIRECTORY
DATA DICTIONARY: DATA INFORMATION FACILITY
FILE INFORMATION
REPORT INFORMATION
PROGRAM INFORMATION
SOURCE INFORMATION
SPECIAL ROUTINE INFORMATION
KEY ELEMENT INFORMATION
PROJECT KEY INFORMATION
ACS WORK ORDER
STANDARD DATA CODE FACILITY
KEY INFORMATION FACILITY
CROSS REFERENCE SYSTEM
CROSS-REFERENCE FACILITY
FILE MATRIX
REPORT MATRIX
PROGRAM MATRIX
INPUT SOURCE MATRIX
DATA SYNONYMS
SYSTEM DATA CODE
EXTEND FILE MATRIX
EXTEND REPORT MATRIX
EXTEND PROGRAM MATRIX
EXTEND INPUT MATRIX

FIGURE 1.

FIGURE 2.

```
                                              ACS DATA ELEMENT DICTIONARY
                                                       06/13/75

  ACS  NO.    DATA ELEMENT NAME           DATA ABBREV NAME          CHAR      UPDATE              DEFINITION
                                                                    CLASS
***
* C3200 TA    CREDIT-CASH                 CASH-CREDIT               N  006    04/07/1975    THE TOTAL AMOUNT OF CASH ADVANCED WHICH WAS R
                                                                                            EPORTED IN THE TRANSMITTAL SHEET 9999.99
* D0100 TA    DATE-ADJUSTMENT             DATE-ADJ                  DATE      04/07/1975    THE DATE MM/DD/YY INDICATES WHEN THE ADJUSTME
                                                                                            NTS WAS MADE
* D0250 TA    DATE-BATCH-TOTAL            DATE-BATCH-TOTAL          DATE      04/07/1975    THE DATE MM/DD/YY INDICATES WHEN THE BATCH TO
                                                                                            TAL WAS UPDATED
* D0350 PM    DATE-BIRTH                  BIRTH-DATE                AN 006    02/15/1975    BIRTH DATE OF THE EMPLOYEE EXPRESSED IN MONTH
                                                                                             DAY AND YEAR
* D0380 CA    DATE-CHANGE                 CHANGE-DATE               DATE      06/04/1975    THE DATE A CHANGE WAS MADE TO ANY ELEMENT BEY
                                                                                            OND ELEMENT NUMBER C19 IN THIS DATA BASE
* D0400 CM    DATE-COMPLETION             COMPL-DATE                DATE      05/22/1975    DATE TRANSACTION WAS COMPLETED
* D0500 PM    DATE-CONTINUOUS-SERVICE     CONT-SER-DATE             AN 006    02/15/1975    DATE FROM WHICH LENGTH OF SERVICE IS CALCULAT
                                                                                            ED FOR EMPLOYEES REHIRED BY JPL
* D0600 TA    DATE-CREDIT                 DATE-CREDIT               DATE      04/09/1975    THE DATE MM/DD/YY INDICATES WHEN THE TRANSMIT
                                                                                            TAL REPORT WAS RECORDED
* D0700 CA    DATE-DELETION               DELETION-DATE             DATE      06/04/1975    THE DATE THE JOB NUMBER IS DELETED
* D0800 PM    DATE-HIRE                   HIRE-DATE                 AN 006    02/15/1975    MOST RECENT DATE AN EMPLOYEE WAS HIRED BY JPL
                                                                                             AND REPORTED FOR WORK
* D1000 CM    DATE-EFFECTIVE              DATE-EFFECTIVE            DATE      05/22/1975    ORIGINAL INSTALLATION DATE
* D1000 PM    DATE-INACTIVE-STATUS        INACTIVE-STATUS-DATE      AN 006    02/15/1975    DATE THAT THE INACTIVE STATUS BEGAN
* D1100 TA    DATE-INVOICE                DATE-INVO                 DATE      04/07/1975    THE DATE MM/DD/YY INDICATES WHEN THE INVOICES
                                                                                             WERE RECORDED
* D1200 PM    DATE-LAST-INCREASE          LAST-INCREASE-DATE        AN 006    02/15/1975    DATE THAT THE MOST RECENT EMPLOYEE RATE INCRE
                                                                                            ASE BECAME EFFECTIVE
* D1400 TA    DATE-RETURN                 DATE-RETURN               DATE      04/07/1975    THE DATE MM/DD/YY INDICATES THE TRAVELER'S RE
                                                                                            TURN
* D1600 CM    DATE-INITIATED              DATE-INITIATED            DATE      05/22/1975    ORIGINATION DATE
```

```
                    CROSS REFERENCE DATA ELEMENT -ALL SYSTEMS
                                    06/12/75

  ACS  NO.   ACS DATA NAME              ACS DATA ABBREV-NAME   PROJ        X-REF
                                                               NAME
***
* E3000 PM  EMPLOYEE-NUMBER         EMP-NO                     TEST1        X
*                                                              PERM4        X
*                                                              TAAS         X
*                                                              PMDBO        X
*                                                              PMFBH        X
* I1100 PM  IND-DEGREE              DEG-IND                    PMFBH        X
*                                                              PMDBO        X
* L3000 CM  LOCATION-EQUIPMENT      EQUIPMENT-LOC              AA35         X
* L4000 PM  LOCATION-EMPLOYMENT     EMP-LOC                    PMDBO        X
*                                                              PMFBH        X
*                                                              PERM4        X
* M0500 PM  MAJOR-BACHELOR          BACH-MAJ                   PMFBH        X
*                                                              PMDBO        X
* N4000 TA  NUMBER-AUTHORIZATICN    AUTH-NO                    TAAS         X
* N4200 CM  NUMBER-BILLING          BILLING-NUM                AA35         X
* N4250 CA  NUMBER-CUSTOMER         CUST-NUM                   COFA         X
* N4260 CA  NUMBER-CUSTOMER-PRCJE   CUST-PROJ-NUM              COFA         X
            CT
* N4270 CA  NUMBER-CUSTOMER-SYSTE   CUST-SYS-NUM               COFA         X
            M
* N4280 CA  NUMBER-CUSTOMER-SUBSY   CUST-SYS-S-NUM             COFA         X
            STEM
* N4290 CA  NUMBER-CUSTOMER-SUE-S   CUST-SYS-SS-NUM            COFA         X
            UBSYSTEM
* N4295 CA  NUMBER-CUSTOMER-SUR-S   CUST-SYS-SSS-NUM           COFA         X
            UB-SUBSYSTEM
* N4300 PM  NUMBER-DIVISION         DIV-NUM                    COFA         X
*                                                              PMDBO        X
* N4400 CM  NUMBER-JPL-ACCOUNT      ACCT-NUM                   AA35         X
```

FIGURE 3.

```
                    DATA ELEMENT PER SYSTEM REPORT MATRIX
                                 06/12/75

  ACS NO.      DATA ABBR. NAME       REPT
                                                               1  1  1  1  1  1
                                   1   2  3  4  5  6  7  8  9  0  1  2  3  4  5
***
* B0100 TA   BATCH-ADJ                                x
* B0200 TA   BATCH-NO                  I              x
* B0400 TA   BATCH-TOTAL               C     C
* B0600 TA   BATCH-TRAV                   I           x  x
* C3000 TA   AIR-CREDIT            x         x        x        x        x
* C3100 TA   BATCH-CREDIT                    I        x
* C3200 TA   CASH-CREDIT           x         x        x        x        x
* D0100 TA   DATE-ADJ                                 x
* D0250 TA   DATE-BATCH-TOTAL                                                x  x
* D0600 TA   DATE-CRED                                x
* D1100 TA   DATE-INVO                                x
*                                                     x
* D1400 TA   DATE-RETURN                              x
* D1800 TA   DATE-TRAV                                x  I
* D4000 TA   DOLLAR-ADJ                               x
* D4100 TA   BALANCE-DOLLAR                           x
* D4300 TA   DOLLAR-INVO           x   x        x     x        x     x  x
* D4900 TA   DOLLAR-TRAV                  x           x  x
* E3000 PM   EMP-NO                x   x     x  x  x  x  x  x  x  x  x  I
* N0500 PM   EMP-FIRST-MID-NAME    x   x  x  x  x  x  x  x  x  x  x  x  x
* N0600 PM   EMP-LAST-NAME         x   x  x  x  x  x  x  x  x  x  x  x  x
* N4000 TA   AUTH-NO
```

FIGURE 4.

173

174

Systems Design Considerations
For the US Army Materiel Command (AMC)
Data Element Dictionary
Directory System

Fernando Puente

US Army Materiel Command
Automated Logistics Management
Systems Agency
St. Louis, Mo.

In early 1967, the Army Materiel Command implemented a Data Element Dictionary system in support of the development of a large complex standard logistics system. The logistics system was designed to manage Army logistics at the wholesale level. The AMC Data Element Dictionary/Directory that evolved has supported system development through its functional and system analysis phases, design, programming, documentation, testing and implementation at six major Army logistics commodity commands located throughout the United States.

Key words: Copy library; data directory; Data Element Dictionary (DED); Central System Design Agency (CSDA); Commodity Command Standard System (CCSS); file guide; Key Word In Context (KWIC); logistics; master file control; programming mnemonics; System Control and Documentation (SYSCAD); system release management.

1. Introduction

The management of military service logistics is by its size, complexity and scope a demanding task. In 1967, the U.S. Army Materiel Command (AMC) began the development of a standard wholesale logistics system that would, when completed, be installed at six commodity commands in the eastern half of the United States. The logistics system that has evolved and is now operational is known as the Commodity Command Standard System (CCSS). The system was developed and fielded by the Automated Logistics Management Systems Agency (ALMSA) located in St. Louis, Mo., which is the Central System Design Agency (CSDA) responsible for its continued maintenance and enhancement. The CCSS includes processes in the functional areas of Financial Management, Provisioning, Cataloging, Stock Control, International Logistics, Equipment Maintenance, Supply Management, and Procurement and Production. It is used to manage at the wholesale level, an inventory of approximately 29 billion dollars. The inventory that is managed ranges from aircraft, missiles, tanks and electronics to ammunition and individual soldier equipment. Throughout the development of the CCSS, a Data Element Dictionary (DED) system has been used as an integral tool supporting its control, documentation, and proliferation.

2. Initial Data Element Dictionary Requirements

In the early planning and development stages of the CCSS design, the challenge and need for data element standardization was recognized, particularly as it related to the integrated storage and file organization principals that were to be used. A single authoritative document for use as a communication link between the system designers, computer systems analyst, programmers, and subject matter specialists was also required. Out of these basic requirements, the AMC Data Element Dictionary/Directory system began to evolve. The first version of the DED system was implemented in early 1967. Its primary purpose was to produce an authoritative document for use by all functional and computer oriented personnel who were designing and programming the CCSS. The dictionary consisted of data element/data field descriptions that included standard names, programming mnemonics, definitions, automated data processing characteristics of the data, and the regulatory reference and authority for the entry. Also included was where used data, as it related to identifiable cells and subcells within the system which equated to major tasks and subtasks which were to be automated within the logistics system. The data elements that were entered in the DED data base were primarily developed by a committee of functionally oriented subject matter specialists, with the assistance of computer systems analysts. Control of the project was exercised by a Systems Integration Division which was charged with coordinating and controlling the total systems design.

3. Dictionary Data Content Development and Usage

The development of the basic data entries for the dictionary involved a great deal of coordination and discussion between the functionally oriented subject matter specialists. Many of the basic data entries that were developed crossed functional area boundaries and in some cases were referred to by differing names or had slightly differing contextual definitions. Before a data element entry was accepted for input into the DED data base, all functional representatives on the committee were required to sign off on the entry as being correct and required by the system. All entries entered into the DED were assigned a functional proponent who was given the responsibility for maintenance after it was entered into the data base. The DED that was produced became the communication link between all the personnel involved in the system development. Because the eventual operational CCSS was to be installed at six different commands in the eastern half of the United States, it also became a very important document to the eventual users of the system as they correlated data in their diverse file to that which would be required for conversion to the standard system master file formats that would be required by the operational system.

4. Dictionary Format

The physical format of the printed DED has remained fairly constant since its first publication. It contains an introductory chapter that explains the content of the dictionary. Next is a Key Word in Context (KWIC) index that contains all words used in the names assigned to the entries in the basic dictionary, listed alphabetically by the occurrence of all the words used in all the names. The KWIC index is used as the primary research tool when trying to determine if a data entry has already been defined in the data base. Next is an alphabetically sequenced mnemonic index cross-referenced to the in the clear names that had been developed for the basic entries. The last section is the basic dictionary, sequenced alphabetically by the names of the data element entries. For the purpose of actually designing systems and for use as a functional definition reference, the sequence for this section has, in our experience, been found to be the most practical when seeking data definitions for either existing systems or for systems that are being designed. This is as opposed to grouping element entries into functional area groupings such as Supply, Cataloging, Procurement, Finance, etc., since so many of the entries cross functional lines. The format also allows for cross-referencing related entries to show hierarchal relationships. The ability to maintain these relationships is designed into the system.

4.1. Individual Entry Content

Every entry in the dictionary contains a unique name that can be a maximum length of 67 character positions. Each entry also contains a unique mnemonic which may be a maximum of 24 character positions and must be constructed as a valid Common Business Oriented Language (COBOL) programming mnemonic. The documentation and programming standards for the CCSS require that this mnemonic be used both when documenting the system or within system application programs which use the COBOL language. Other required attributes of the data element entry are the characteristics of length, type, and data field justification, the regulatory authority or reference for the entry, data field security, the date the entry was last changed, the identity of the proponent, the definition, what cells and sub-cells the element is used in, and if the element is locally derived, all its data codes and items. Also provided is the capability to register other known references to uses of the entry related to local publication references, higher level regulatory references, report usages related to Report Identification Numbers (RIN's), Report Control Symbols (RCS's) and forms usage. Application programs, file usage, systems usage references and known synonym mnemonic/abbreviations may also be registered. All of these references are available on automated cross-reference indexes for use in as required inquiry processes.

5. Programming Mnemonic Usage

In the first design of the data element dictionary system, all mnemonics assigned to data element entries were limited to a maximum of 15 character positions. Through experience, this was found to be unsatisfactory, particularly when attempting to create meaningful programming mnemonics when documenting long names and attempting to maintain a semblance of standard abbreviations within the mnemonic. In a subsequent redesign of the data element dictionary system, the allowable length of a mnemonic was increased to 30 character positions to match the full capability of the COBOL language. In the CCSS, the maximum allowable length was increased to 24 character positions. Six positions are reserved to identify data fields within the hierarchally structured master files of the CCSS. Since the CCSS was to be programmed in COBOL, it was recognized that an important tool for managing the system would be the mnemonic programming tags that would be used within the application programs and would also be used to describe the data fields within the CCSS master files. Because application programs written against system master files would be required to use standard file descriptions which were resident on the system libraries, a great degree of discipline was imposed on the application programmer as he manipulated data within the master files or passed data to other application queues for use in other system processes. Because the programming mnemonic was to be embedded within the application programs and the master file descriptions, and to a certain degree reflect what was being manipulated within the system, the mnemonic was made the major key of the data element dictionary system, which in turn dictated that all mnemonics developed for entry into the data element dictionary data base must be unique.

6. Multiple Remote User Capability

Early in the dictionary system development, a decision was made to allow other Army Materiel Command system design activities to register as users of published dictionary entries or, if necessary, develop and become proponent for entries that had not been previously entered. This requirement led to the development of the capability to accept input from diverse activities and register their interest to entries in the dictionary, or accept submittal of new entries. Also developed was the capability to protect proponents of data entries in the dictionary from unauthorized update of his data. This was accomplished in the edit processes by developing system tables that cross-checked submitter input codes to entry proponent identifications in the data base. The DED data base is updated in a batch mode on a monthly basis. Preparatory input data edit cycles are run on an as required basis. After update, two copies of the dictionary are printed for use by system operation control personnel. It is also reproduced on microfilm for reference by inhouse and remote activity personnel. The microfilm contains all basic dictionary listings plus audit trail listings which include editing process results and before and after update images for all data base entries that were affected in the update.

7. Master File Directory Development

To control and document the CCSS master files, a companion data base to the DED system was developed to act as a directory for the data that was being stored in the various master files of the CCSS. This portion of the system when fully developed was named the Systems Control and Documentation (SYSCAD) system. By coupling the two data bases, data element entries which were to be in the CCSS master files could be edited against the DED data base to insure that mnemonics that were being used in the CCSS master files descriptions had been established and that the automated data processing field characteristics were compatible with the DED entries. Since the SYSCAD master file contained all the CCSS master file descriptions, standard COBOL copy descriptions in both COBOL F and ANS COBOL were generated and established on the systems libraries for mandatory use of the application programmers when accessing any CCSS master file. The SYSCAD system when coupled to the DED data base also has the capability to generate tailored file descriptions that contain definitions for all data fields in a CCSS master file plus the structure of the file. This product documents the master files and is disseminated as a file guide publication to the users of the CCSS. The file guide publications are considered as the single source, official documentation of the content and structure of the master files.

7.1. Master File Directory Usages

As mentioned previously, the CCSS was designed to be installed as a standard wholesale logistics system at six sites throughout the eastern United States. Well in advance of the installation of the CCSS, the eventual users had to know the structure and content of the CCSS master files so that they could begin correlating data from their system files to that which would be required for the standard CCSS file structures and data content. The file guide publications produced by the SYSCAD system were the primary documents used for this purpose. Also provided the user from the SYSCAD system, for use as an automated file conversion aid, were magnetic tapes that contained element by element descriptions of the master files for use in automated conversion methodologies. These magnetic tapes also contained change indicators so that the user could be kept informed of any changes that might be occurring to the master files. Also contained in the SYSCAD system are computed data starting positions for all data fields identified in the CCSS master file and computed segment lengths for both fixed and variable length records. The descriptions also identify all data fields that are used as file keys. Operationally, length and key position information is also used for keeping internal CCSS software control tables in synchronization.

7.2. Master File Version Control

The CCSS in its present configuration contains 29 master files. For the most part, files that were developed were designed to reflect the data required by the various functional logistics areas, taking into consideration the conservation of processing time, effective use of storage devices, ease of data maintenance and protection of data integrity within the system. These files in certain run configurations may stand alone or be coupled with other master files to run in a data sharing configuration. The CCSS has one primary file that contains common data that may be used selectively or in combination by any functional process in the CCSS. All other files may be influenced directly or indirectly by this file. The structure of these files is controlled and documented through the use of the DED system and SYSCAD system. Since the first phase of the system became operational in April 1971 at the U.S. Army Aviation Systems Command in St. Louis, Mo., the CCSS has continued to be changed and enhanced. This, of course, has included changes to the master files structure and content under a concept of scheduled system release management. In support of the release management concept, the SYSCAD system is designed to maintain, control, and contain previous master file structures for regression test purposes, current file structures and future file structures as they will be in projected releases of the system. With the capability to manage future file structure releases, standard file description can be made available to the application programmer with sufficient lead time so that system changes may be coded and tested for whatever system release that the changes are required. This capability is very important because the magnitude and scope of a future systems change may require that work begin on a systems change many months before it is actually fielded.

8. Summary

The AMC data element dictionary/directory data base presently contains approximately 8000 data field definitions which have been identified to logistic processes within automated systems. The system has continued to evolve through a need for communication, system standardization, documentation, and control. It is only one answer to what can be done to support and attempt to control automated systems. The true benefits from establishing a Data Element Dictionary/Directory system is when the data that has been collected or is being collected, supports system developers by making their task easier and allows them to produce more efficient, well documented and controlled systems.

A Data Element Directory for a State Motor Vehicles Agency

John Roberts

Data Standards and Controls Bureau
New York State Department of Motor Vehicles

The New York State Department of Motor Vehicles administers one of the most comprehensive data processing installations in the nation. The socially and economically significant data in its computer files is a resource demanding protection and management.

Data elements are identified, defined and their representations documented in the Department of Motor Vehicles Data Element Directory (DMV/DED). This directory is the central reference for the agency's data resources.

This paper describes the content, organization and methodology of the DMV/DED.

Key words:[1] data element; Data Element Descriptions; data element directory; definition information; Explanatory Text; Files Inventory; identification information; Transactions Inventory; Word Reference

1. Background

The New York State Department of Motor Vehicles administers the State's Vehicle and Traffic Law. Major agency programs are:

(a) Licensing and controlling motor vehicle operators

(b) Registering and titling motor vehicles

(c) Conducting research and education in traffic safety areas

(d) Adjudicating traffic violations in New York City, Buffalo and Rochester

(e) Requiring proof of financial security coverage for motor vehicles

---

[1] A glossary of terms associated with the DMV/DED is appended to this paper

These programs require the Department of Motor Vehicles to maintain large quantities of data for users. An overview of the data maintained and users involved was presented at the first symposium by Emswiler and Heitzler. (1)[2]

> "Each of the states maintains records required for the registration, licensing and control of motor vehicles and drivers. These data systems, 49 of which are automated in whole or in part, make up one of the largest sets of data files in the nation. They perform the primary production functions of revenue collection, vehicle/driver identification and control, and motor vehicle code administration of the individual states. These functions require the retention of motorist data and the interchange of that data between state agencies, individual states and other users of the systems. These systems have been tapped by law enforcement, by commercial users and by some federal systems. As much as 70 percent of law enforcement data traffic is driver/vehicle related. Commercial users are insurance, credit, employer and statistical collection interests. The Federal systems include law enforcement and traffic safety."

Federal, New York State and local law enforcement agencies are directly connected to the Department's computer system. Most New York State County Clerks act as agents for the Department of Motor Vehicles on a statutory fee basis. County motor vehicle issuing offices are essential links in the Department's statewide, teleprocessing network.

The heart of the Department's computer system is two IBM 360/65 central processing units operating in tandem and providing multi-processing capability. Over 675 terminals (326 buffered hard copy and 350 video display) enter and retrieve data from the system. Additional, direct computer to computer interfaces with law enforcement agencies provide inquiry capability for an additional 400 video displays.

---

[2] Figures in parentheses refer to literature references at the end of the paper

2. Acknowledgement

The content and data collection tools for the Department of Motor Vehicles Data Element Directory (DMV/DED) are based upon material developed by the ANSI States' Model Motorist Data Base, Data Directory Committee (D-20.1). Specific materials used were the D-20.1 Committee's data collection form and related instructions. (1)

3. Data Resources

The main data resources of the Department of Motor Vehicles are its computer master record files. The first illustration identifies these files and describes their content.

NEW YORK STATE DEPARTMENT OF MOTOR VEHICLES
DATA ELEMENT DIRECTORY (DMV/DED)
MAJOR COMPUTER MASTER RECORD FILES

FILE NAME	NUMBER OF RECORDS	NUMBER OF DATA ELEMENTS
LICENSE FILE · Administrative and Historical Driver Licensing Information	12,377,598	45
REGISTRATION & PLATE INDEX FILES · Vehicle Registration and Plate Inventory Information	REG. PLATE 22,625,266	10
	REG. 14,299,056	54
VIN TITLE · Motor Vehicles Ownership and Lien Information	8,594,428	52
ACCIDENT RECORDS FILE Motor Vehicle Accident Cases	864,060	63

First Illustration

The Division of Data Administration was organized in 1972 to coordinate the management and protection of the Department's data resources. The Data Standards and Controls Bureau, a component of the Data Administration Division, was directed in 1974 to develop and maintain a Department of Motor Vehicles Data Element Directory (DMV/DED). This directory is intended to be a central data reference guide.

The DMV/DED identifies and defines Departmental data elements and documents their input, file and output representations. The directory links data elements to records, files and transactions. The production of the DMV/DED will not require changes to existing information structures but it will be used to standardize data. A description of this type of data element directory was developed by Bontempo and Swanz in their article "Data Resources Management". (2)

> "The DED is thus a single, authorative source of information on data elements, their organization and format. It is a way of monitoring and controlling data resources without actually integrating and centralizing the data itself."

4. Content

The definition of a data element contained in the glossary of this paper is based upon material presented in the ANSI Data Standardization Criteria Task Group's (X3L81) proposed guide for developing data representation standards. (5) The data element is the fundamental building block from which information structures are made.

The data element consists of a general part which designates the information required and a specific part called the data item which supplies the information required. For example, the general part "Color of Hair" can have several data items or values, "Black", "Brown", etc.

The DMV/DED contains three categories of information: identification, definition and representation.

Identification information furnishes the data element with unique labels and links the element to records, files and transactions. The identification information category contains items which associate DMV/DED data elements with those in the directories of external users.

Definition information explains the content, purpose, source and use of the data element.

Representation information describes the length, type and format of the data item or specific part of the data element.

The second illustration lists the three information categories and the items contained in each category. A full explanation of each category item is contained in the glossary of this paper.

NEW YORK STATE DEPARTMENT OF MOTOR VEHICLES
DATA ELEMENT DIRECTORY (DMV/DED)

INFORMATION CATEGORIES

IDENTIFICATION

Abbreviation
DMV Code
Files (Including Records)
Name-In-Context
NHTSA Reference
Signature
Synonyms
Systems
Traffic Records Reference
Transactions

DEFINITION

Additional Comments
Input Source
Special EDP Use
Technical Definition
Type of Data Element
Use

REPRESENTATION

Data Item Description
Format of Representation
Length of Representation
Level of Representation
Type of Representation

Second Illustration

5. Organization

The DMV/DED consists of five sections:

(a) Explanatory Text

(b) Data Element Descriptions

(c) Files Inventory

(d) Transactions Inventory

(e) Word Reference

5.1 Explanatory Text

The Explanatory Text for the DMV/DED is developed from the directory project work plan and design specifications for the four other sections. This section describes the organization, purpose and maintenance of the directory. All terms used in the DMV/DED are defined in the Explanatory Text.

5.2 Data Element Descriptions

This section contains a complete description for each computer processed data element in the Department of Motor Vehicles. All information categories and their items are documented. The third illustration outlines the format for a Data Element Description. Data elements are arranged in alphabetical order by signature.

NEW YORK STATE DEPARTMENT OF MOTOR VEHICLES
DATA ELEMENT DIRECTORY (DMV/DED)
DATA ELEMENT DESCRIPTION

SIGNATURE AND DMV CODE:	Date of Birth (015C).
NAME–IN–CONTEXT:	"Date of Birth" on License Certificate Form.
ABBREVIATION:	D.O.B.
TECHNICAL DEFINITION:	The day of one's birth expressed as month, day and year.
INPUT SOURCE:	Applicant for a Driver's License and/or Registration Certificate.
USE:	Identification and Control.
SYNONYMS:	Birthday, Day of birth.
COMPUTER FILES:	Title (Activity-Ownership Trailers)
TRANSACTIONS:	This element is used in all License and Registration Transactions.
SYSTEMS:	License, Registration, Title.
COMPONENTS:	151B Year; 152B Month, 153B Day.
SPECIAL EDP USE:	This element is used as part of an access key to the License and Registration files.
COMMENTS:	This element is included as part of the Motorist Identification Number on the License and Registration Files.
TRAFFIC RECORDS REFERENCE:	01. 01 Date of Birth 2.0.00H
NHTSA REFERENCE:	2.0.00H Date of Birth

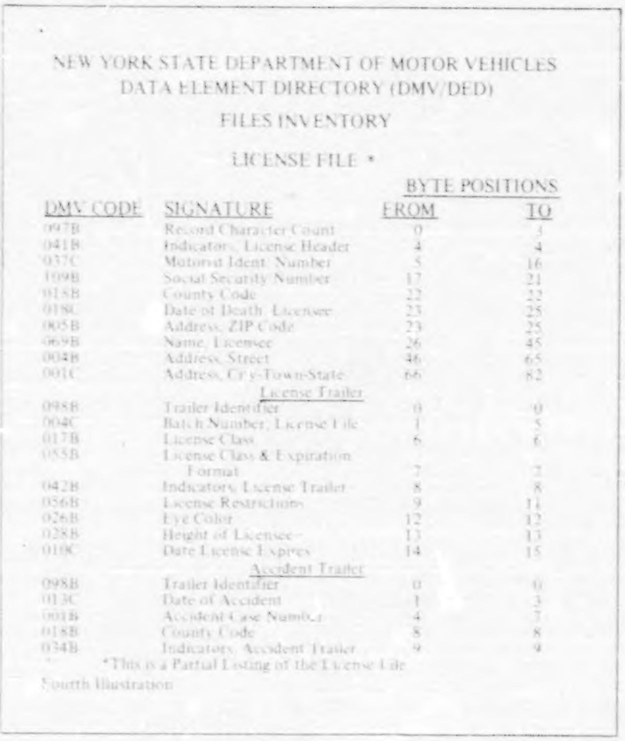
REPRESENTATIONS:

LEVEL	TYPE	LENGTH	FORMAT	DATA ITEM DESCRIPTION	
INPUT	CODE	6 POS.	NUMERIC NO SIGN	MM DD YY	MM=01 to 12 DD=01 to 31 YY=00 to 99
FILE	CODE	3 BYTES	PACKED DECIMAL NO SIGN	YY MM DD	SAME VALUES AS INPUT
OUTPUT (CRT/PRINT)	CODE	8 POS.	NUMERIC NO SIGN	MM/DD/YY	SAME VALUES AS INPUT

Third Illustration

5.3 Files Inventory

The Files Inventory associates data elements with the Department's computer master record files and their records. The master record files are arranged in alphabetical order by file name. The format for the Files Inventory is shown in the fourth illustration. Each file's records are listed. The data elements in each record are identified by DMV code, signature and record byte positions.



NEW YORK STATE DEPARTMENT OF MOTOR VEHICLES
DATA ELEMENT DIRECTORY (DMV/DED)

FILES INVENTORY

LICENSE FILE *

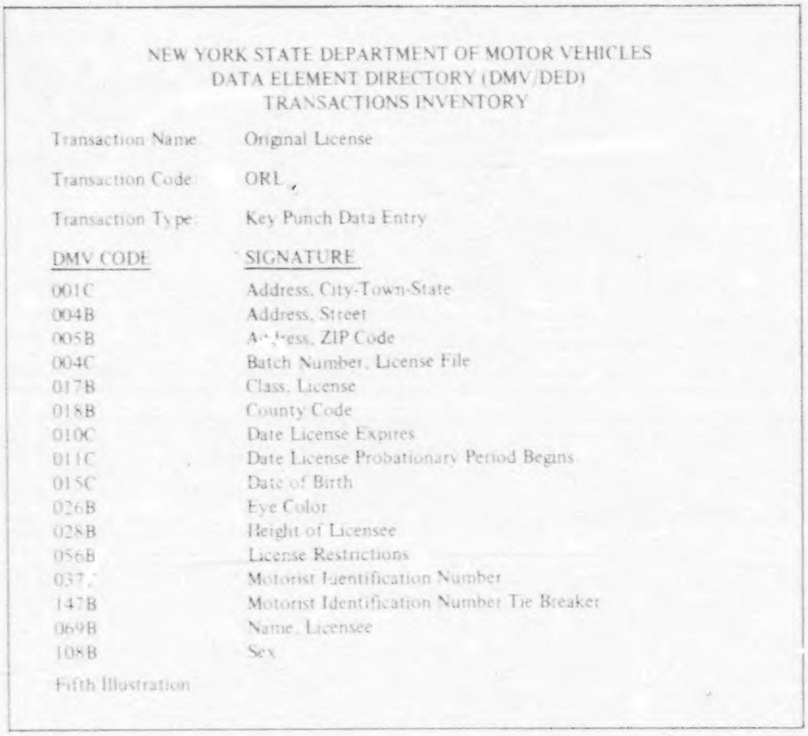
		BYTE POSITIONS	
DMV CODE	SIGNATURE	FROM	TO
097B	Record Character Count	0	3
041B	Indicators, License Header	4	4
037C	Motorist Ident. Number	5	16
109B	Social Security Number	17	21
015B	County Code	22	22
018C	Date of Death, Licensee	23	25
005B	Address, ZIP Code	23	25
069B	Name, Licensee	26	45
004B	Address, Street	46	65
001C	Address, City-Town-State	66	82
	License Trailer		
098B	Trailer Identifier	0	0
004C	Batch Number, License File	1	5
017B	License Class	6	6
055B	License Class & Expiration Format	7	7
042B	Indicators, License Trailer	8	8
056B	License Restrictions	9	11
026B	Eye Color	12	12
028B	Height of Licensee	13	13
010C	Date License Expires	14	15
	Accident Trailer		
098B	Trailer Identifier	0	0
013C	Date of Accident	1	3
001B	Accident Case Number	4	7
018B	County Code	8	8
034B	Indicators, Accident Trailer	9	9

*This is a Partial Listing of the License File

Fourth Illustration

5.4 Transactions Inventory

The Transactions Inventory links data elements with the Department's computer transactions. Transactions are identified by transaction name, reference code and type. Each transaction's data elements are listed in alphabetical order and identified by DMV code and signature. The format for the Transactions Inventory is shown in the fifth illustration.



NEW YORK STATE DEPARTMENT OF MOTOR VEHICLES
DATA ELEMENT DIRECTORY (DMV/DED)
TRANSACTIONS INVENTORY

Transaction Name: Original License

Transaction Code: ORL

Transaction Type: Key Punch Data Entry

DMV CODE	SIGNATURE
001C	Address, City-Town-State
004B	Address, Street
005B	Address, ZIP Code
004C	Batch Number, License File
017B	Class, License
018B	County Code
010C	Date License Expires
011C	Date License Probationary Period Begins
015C	Date of Birth
026B	Eye Color
028B	Height of Licensee
056B	License Restrictions
037[illegible]	Motorist Identification Number
147B	Motorist Identification Number Tie Breaker
069B	Name, Licensee
108B	Sex

Fifth Illustration

5.5 Word Reference

This section identifies the use of all words (signatures, names-in-context, abbreviations, synonyms and components) in the Data Element Descriptions section of the DMV/DED. The following information is presented for each word:

1. DMV code and signature of the element where the word appears
2. Specific use of the word in the data element (e.g. signature, synonym, abbreviation, etc.)

Words are listed alphabetically. Some words are used in several data elements. These words have multiple listings in the Word Reference. The sixth illustration indicates the Word Reference format.

NEW YORK STATE DEPARTMENT OF MOTOR VEHICLES
DATA ELEMENT DIRECTORY (DMV/DED)
WORD REFERENCE

Word	Appears As	In Data Element	
		DMV Code	Signature
Fuel	Synonym	122B	Type of Fuel
Month	Signature	152B	Month
Month	Component	015C	Date of Birth
Name	Name-In-Context	069B	Name, Licensee
Name, Licensee	Signature	069B	Name, Licensee
VIN	Abbreviation	038C	Vehicle Identification Number

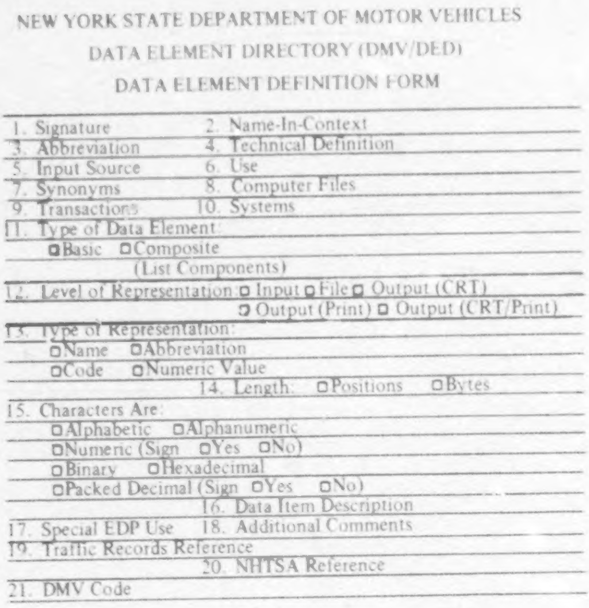
Sixth Illustration

6. Data Collection Tools

The data collection tools for the DMV/DED are:

1. Data Element Definition Form - used to record identification and definition information as well as file representation information.
2. Data Element Representation Form - used to record input and output representation information.

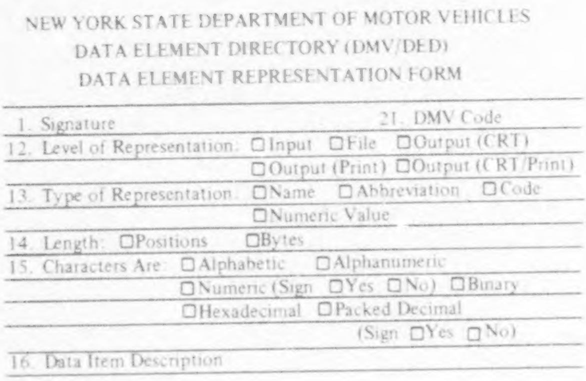
The seventh and eighth illustrations are condensed versions of these forms.



NEW YORK STATE DEPARTMENT OF MOTOR VEHICLES
DATA ELEMENT DIRECTORY (DMV/DED)
DATA ELEMENT DEFINITION FORM

1. Signature 2. Name-In-Context
3. Abbreviation 4. Technical Definition
5. Input Source 6. Use
7. Synonyms 8. Computer Files
9. Transactions 10. Systems
11. Type of Data Element:
☐Basic ☐Composite
(List Components)
12. Level of Representation: ☐ Input ☐ File ☐ Output (CRT)
☐ Output (Print) ☐ Output (CRT/Print)
13. Type of Representation:
☐Name ☐Abbreviation
☐Code ☐Numeric Value
14. Length: ☐Positions ☐Bytes
15. Characters Are:
☐Alphabetic ☐Alphanumeric
☐Numeric (Sign ☐Yes ☐No)
☐Binary ☐Hexadecimal
☐Packed Decimal (Sign ☐Yes ☐No)
16. Data Item Description
17. Special EDP Use 18. Additional Comments
19. Traffic Records Reference
20. NHTSA Reference
21. DMV Code

Seventh Illustration



NEW YORK STATE DEPARTMENT OF MOTOR VEHICLES
DATA ELEMENT DIRECTORY (DMV/DED)
DATA ELEMENT REPRESENTATION FORM

1. Signature 21. DMV Code
12. Level of Representation: ☐Input ☐File ☐Output (CRT)
☐Output (Print) ☐Output (CRT/Print)
13. Type of Representation: ☐Name ☐Abbreviation ☐Code
☐Numeric Value
14. Length: ☐Positions ☐Bytes
15. Characters Are: ☐Alphabetic ☐Alphanumeric
☐Numeric (Sign ☐Yes ☐No) ☐Binary
☐Hexadecimal ☐Packed Decimal
(Sign ☐Yes ☐No)
16. Data Item Description

Eighth Illustration

The Data Element Definition Form is letter size with items printed on both sides of the form. The Data Element Representation Form is one-third the depth of the definition form and is printed three forms to a letter size page.

During the production of the DMV/DED these forms are the main repository for data element description information. After the DMV/DED is published, the forms will be used to amend present descriptions and obtain information about new data elements.

7. Methodology

Most projects have these common steps:

(a) Define objectives

(b) Obtain management approval

(c) Research literature

(d) Acquire and train project staff

(e) If a computer project, arrange for systems and programming help

This methodology assumes these steps and concentrates on tasks essential to producing the DMV/DED.

7.1 Content Determination

Determining what should appear in the publication is a critical decision in the development of the data element directory. Non-essential items will overload the directory. Omission of important items will diminish the directory's usefulness. Content determination is derived from an analysis of the needs of data processing professionals, managers, external and internal users. This analysis should be supplemented by a review of existing data element directories and available literature.

7.2 Design Specifications

The directory data collection forms and output formats must be designed to include all content items. All items appearing on forms and in the directory have to be clearly defined.

An important factor in the design of the output specifications is the method used to publish the directory. The DMV/DED is presently typewritten on letter size paper and stored in loose leaf binders. When the DMV/DED project is completed in late 1975, the publication will be stored on a direct access device and reports generated on computer printout paper. The output specifications for the DMV/DED were developed to accommodate both methods of publication.

7.3 Maintenance and Control

Data is an evolving resource. The data element directory must be inherently able to be amended and updated to reflect the data it describes. Systems and procedures for maintenance and change have to be developed before the project begins.

Changes to the directory must be carefully controlled. Responsibility for altering the directory should be given to the agency's data base administrator.

7.4 Explanatory Text

The Explanatory Text should be written from the project work plan; content items; forms and output specifications and related instructions. This text is useful as a tool for training the project staff and familiarizing managers and users with the directory project.

7.5 Files Inventory

The Files Inventory is concerned with data elements on the agency's computer master record files. These files contain the most significant data elements. Information for the Files Inventory is derived from systems and programming documentation. If documentation is inadequate, interviews with data processing professionals and users must be scheduled.

7.6 Transactions Inventory

The Files Inventory will identify data elements on the master record files. The Transactions will describe how the elements get to and from the files. This inventory will also identify data elements not found on the files. Input and output representation information is derived from the Transactions Inventory.

Interviews with data processing professionals, managers and users are the most reliable sources of Transactions Inventory information. When conducting interviews, all forms, reports and documentation associated with the transactions should be collected. These items will supply data about identification and definition information.

7.7 Data Element Descriptions

After the Files and Transactions Inventories are completed, all information from these inventories should be transferred to the data collection forms for the data elements. When this step is accomplished, the data collection forms should be reviewed to determine what additional information must be collected. A work plan should be developed for gathering this data. Information may have to be obtained from sources outside of the agency (e.g. Traffic Records Reference data was obtained from the New York State Traffic Records Project).

The Data Element Descriptions can be prepared for clearance when the additional information has been collected. The preparation of descriptions from the data collection tools is a task requiring accuracy and control.

The Data Element Descriptions provide the words (signatures, names-in-context, etc.) for the Word Reference. This directory section and the Data Element Descriptions should be completed simultaneously.

7.8 Clearance and Publication

The data element directory should be cleared by the managers, users and data processing professionals who will use the publication. Sufficient time should be alloted for a thorough review. All significant comments should be considered before publication.

After clearance, the publication and maintenance of the directory should be the responsibility of the agency's data base administrator.

8. Conclusion

The DMV/DED will be used to standardize data in the Department of Motor Vehicles. A goal of this standardization will be to promote data interchange among organizations involved in traffic safety.

Other state motor vehicles agencies are encouraged to obtain further information about the DMV/DED.

9. References

1. C. E. Emswiler, Jr. and C. P. Heitzler, Jr., "The States' Model Motorist Data Base Project American National Standards Institute (ANSI) D-20" Management of Data Elements in Information Processing. (Proceedings of a Symposium sponsored by the American National Standards Institute Committee X3L8 and the National Bureau of Standards). U. S. Department of Commerce, National Technical Information Service, CBM 74-10700, April 1974.

2. American National Standards Institute (ANSI) Committee X3L81, "Guide for the Development, Implementation and Maintenance of Standards for the Representation of Computer Processed Data Elements", Proposed Technical Report X3L8/186, 73-08-01 (This report appeared as Appendix B in the Management of Data Elements In Information Processing.)

3. Charles J. Bontempo and Donald G. Swanz, "Data Resources Management", Data Management February 1973.

10. Glossary

Abbreviation	The identification category item that indicates the shortened form of the data element signature used on reports.
Additional Comments	The definition category item used to describe information about the data element not covered in other information category items.
Basic Data Element	A data element that conveys a singular object (e,g. "color of eyes" connotes "blue").
Composite Data Element	A data element that conveys information whereby multiple facts can be derived. The parts that make up the composite data element are referred to as components. In the example given, the components of the data element "Date of Birth" are: 1. Year of Birth 2. Month of Birth 3. Day of Month of Birth Components are basic data elements in the DMV/DED.
Data Element	A unit of meaning used to identify the field in a record. The data element is the fundamental building block from which information structures (records, files, data bases) are derived. The data element consists of two parts: 1. A general part which designates the information required. 2. The data item which supplies the information required.
Data Element Definition Form	The basic data collection tool for the data element directory.
Data Element Descriptions	The data element directory section containing complete identification, definition and representation information about individual data elements.

Data Element Directory	The central data reference that identifies and defines an agency's data elements and documents their representations. The directory links elements to the agency's information structures (records, files, data base).
Data Element Representation Form	The data collection tool used to record representation information. This form is derived from the Data Element Definition Form.
Data Item Description	The representation category item that describes and illustrates the values or representations of the data element.
Definition Information	The information category that describes the content, purpose, source and use of the data element.
DMV Code	The identification category item used to record the Department of Motor Vehicles code assigned to the data element for ease in manipulating data.
DMV/DED	The abbreviated title of the New York State Department of Motor Vehicles Data Element Directory.
Explanatory Text	The data element directory section that describes the content, purpose and use of the directory.
Files	The identification category item that lists the computer master files where the data element appears.
Files Inventory	The data element directory section that identifies the agency's computer master record files and their data elements.
Format of Representation	The representation category item that designates the characters (bytes or positions) used to represent the data element (e.g. alphabetic, alphanumeric, binary, numeric, etc.).
Header/Trailer Record	A collection of related data elements treated as a unit. Header records contain the name and address of an individual. Trailer records contain historical and/or administrative information about an individual (e.g. accident trailer; registration certificate trailer; lienholder trailer).

Input Source	The definition category item that denotes the originator of the data element.
Length of Representation	The representation category item that indicates the number of positions or bytes of the values in the data item descriptions.
Level of Representation	The representation category item that designates the level of processing (input, file, output) being described in the data item description.
Name-In-Context	The identification category item that describes how the signature of the data element appears on forms where uniqueness is provided by context. (e.g. "Name, Licensee" appears as "Name" on the License Certificate).
NHTSA Reference	The identification item used to link DMV/DED data elements to entries in the National Highway Traffic Safety Administration's "Design Manual for State Traffic Records Systems".
Representation Information	The information category that describes the length, format and characteristics of the data item, the specific part of the data element.
Signature	The identification category item that denotes the complete name of the data element.
Special EDP Use	The definition category item that describes the data element's use for specific data processing purposes (e.g. access key, searching field).
Synonyms	The identification category item that lists names or words similar or identical in meaning to the signature of the data element.
Systems	The identification category item that describes the major computer procedures that process the data element.

Technical Definition	The definition category item that explains the content and purpose of the data element.
Traffic Records Reference	The identification category item that links DMV/DED data elements to data elements in the data element directory of the New York State Traffic Records Project. The New York Traffic Records Project is a federally funded agency charged with improving the means for reporting, processing and analyzing traffic records data.
Transactions	The identification category item that designates the computer transactions that use the data element.
Transactions Inventory	The data element directory section that identifies the agency's computer transactions and their data elements.
Type of Element	The definition category item that classifies the data element as basic or composite. (See separate glossary entries for basic and composite data elements).
Type of Representation	The representation category item that specifies the method in which the values in the data item description are recorded. There are four representation types: 1. Name - a word that conveys the meaning of the data element. 2. Abbreviation - a shortened form of the name of the data item. 3. Code - a fixed length representation that describes the data element. 4. Numeric value - A representation that conveys mathematical or measurement meaning.
Use	The definition category item that denotes the general uses of the data element.
Word Reference	The data element directory section that describes how words (signatures, names-in-context, abbreviations, synonyms and components) are used in the Data Element Descriptions.

A DATA ELEMENT DIRECTORY
FOR
A STATE MOTOR VEHICLES AGENCY

ADDENDUM

INTRODUCTION

The questions about this paper raised at the second data element management symposium can be grouped into four broad categories:

A. Maintenance and control of the DMV/DED.

B. Resolution of identification, definition and representation conflicts.

C. Restriction of data element information.

D. Adaptation of FIPS and/or ANSI Standards.

This addendum briefly describes each of these areas.

MAINTENANCE AND CONTROL OF THE DMV/DED

The DMV/DED is being developed in a non-data base environment as a central data reference. After clearance and publication, the directory will be used to standardize data element identification, definition and representation information.

The development and standardization process will require the directory to be continually revised. The Data Administration Division staff will be responsible for acquiring information to maintain the directory's accuracy and completeness.

Major sources of change information include:

1. Systems and programming documentation.
2. Policy and system planning papers.
3. Systems design packages and modifications.

Data Administration staff will be actively soliciting and obtaining information from these sources while the directory is being developed. After publication, procedures will be installed to ensure that change information is received and reviewed by the Data Administration Staff.

Control of the content of the directory will follow established control procedures of the Department of Motor Vehicles.

These procedures include the following steps:

1. Obtain full information about the change and why it is being made.
2. Analyze the effects of the change.
3. Prepare a written report of the change and its effects.
4. Clear the report with relevant operating managers, data processing professionals and outside users.

5. Incorporate comments obtained during the clearance process in the final report.

6. Make necessary changes through the unit responsible for maintaining the directory. The Department's data base administrator or the Directory of Data Administration is responsible for maintaining the DMV/DED.

The development of the DMV/DED in a non-data base environment facilitates the management and change of the directory during development. Since the directory is not on-line, manual changes can be fully controlled and implemented by non-EDP staff members.

When the Department organizes a data base, the DMV/DED would be expanded to include full logical and physical data. It would then be a part of the data base management system. The maintenance and control procedures would have to be revised to reflect the level of sophistication, the needs of the data base management system and the working rules for the data base environment.

RESOLVING CONFLICTS

The identification of information conflicts is a primary reason for developing the DMV/DED. The resolution of conflict is a major reason for standardizing data elements. The major step in conflict resolution will be in recognizing that a conflict exists. Once identified, the definition of the conflicting areas and suggestions to resolve the conflict will be done as a part of the standardization of data elements.

RESTRICTION OF DATA ELEMENT INFORMATION

The Department of Motor Vehicles is developing an information disclosure policy to guard the privacy and rights of individuals identified in its data base. The main provisions of this policy include classifying data elements into various information categories. This classification also covers how the data elements are used in transactions and as they appear on files. The categories selected are:

Administrative - This includes all data that is to be used in the Department of Motor Vehicles but which is not to be given out to law enforcement agencies, other public agencies or private individuals or groups.

Law Enforcement Data - This data is information that is used in the Department of Motor Vehicles and is given out only to law enforcement agencies. This information is never given out to non-law enforcement public agencies and/or private individuals or groups.

Public Information - This data is information that is used in the Department and can be given out to law enforcement agencies, non-law enforcement public agencies and private individuals or groups.

When implemented, this policy will be shown in the data element directory. Each data element will contain a restriction classification. These classifications will also apply to transactions and the files inventories.

The implementation of this policy will require many changes to the present systems and programming of the Department's application areas.

ADAPTATION OF FIPS OR ANSI STANDARDS

After the publication of the DMV/DED the standardization effort will include as a basic step the review of all relevant standards developed by the ANSI and FIPS Committees. This review will determine the schedule for standards implementation. As explained in the written presentation, the DMV/DED is already based upon standards developed by two ANSI committees. The data collection form and instructions of the ANSI D-20 Committee State Model Motorist Data Base are the nuclei for the DMV/DED. Standards suggested by the ANSI X3 Committee on Data Element Representation were used in developing techniques for the directory.

An Integrated Dictionary for Systems and Data Components

Curg Shields[1]

M. Bryce and Associates, Inc.
1248 Springfield Pike
Cincinnati, Ohio 45215

This paper discusses two products developed and marketed by M. Bryce & Associates, Inc. "PRIDE", PRofitable Information by DEsign, is a planned approach to Information Systems design, development and implementation currently installed in over 430 systems organizations. "PRIDE"-Logik, Logical Organizing and Gathering of Information Knowledge, is an automated systems and data dictionary for use with "PRIDE". Both are proprietary products, copyrighted and trademarked by M. Bryce & Associates, Inc.

Included in this paper will be a general discussion of these products and user reactions to an integrated systems and data dictionary.

Key words: Data; data dictionary; data management; information; information systems; methodology; systems dictionary.

M. Bryce & Associates, Inc. is a management consulting firm specializing in the field of Information Systems. During the past four years, our major emphasis has been the marketing and support of an Information Systems design methodology called "PRIDE". "PRIDE" includes coverage of the design method, project management, data management and documentation as part of Information Systems development. The methodology contains all the activities required to design, develop and implement an Information System. Project Management covers the estimating, scheduling and controlling of an information system design project. Data Management identifies, monitors and controls data from its source through its end use as information. Documentation provides both working and ongoing documentation prepared during the design process.

"PRIDE's" design philosophy is based on the fact that all Information Systems have a consistent structure regardless of their size or function.

---

[1] Vice President, Technical Support

Briefly stated, systems contain one or more sub-systems. Sub-systems consist of one or more procedures, some manual and possibly some computer. Manual procedures contain operational steps executed by one or more people. Computer procedures include one or more programs executed by a computer.

"PRIDE's" data management concept forces the analyst to first define the information requirements of the system's end users. Next the analyst identifies the data necessary to meet the information requirements. This concept makes the assumption that we as professionals should know the difference between information and data. This may be an incorrect assumption since much of our trade literature has a tendency to use these terms interchangeably. Because a significant and operable difference does exist, I will define these terms.

DATA is a digital representation of a fact or event. It can be Indicative, Descriptive or Quantitative. Data only has meaning when related to other data.

INFORMATION is created as a result of collecting, processing and analyzing Data. Information provides the user with the facts required to take action or make a decision. The accuracy of Information depends upon the validity and completeness of the data and the processing logic used. Information, therefore, is the understanding or insight gained from the processing and analysis of data.

When the data requirements for a system are determined then the analyst designs the inputs to collect the data, the outputs to report the information and the files to store the data. Finally the necessary processing steps for the system are defined.

When designing an Information System, it is recognized that the analyst must consider data contained in manual files as well as computer files. This Data Management concept is not limited to the traditional, narrow viewpoint of computer data bases and data base administration. It includes all data regardless of where stored or how processed.

Applying both the system structure and data management concepts allows the systems designers to identify the system and data component requirements of an Information System. Both concepts also establish the relationships among the systems components, among the data components and between the systems and data components with particular attention to the 'where used' trails. When following this approach, for designing and documenting information systems, our "PRIDE" Users began to realize that the manual documentation could eventually become rather difficult to maintain, especially with the multiplicity of interrelationships established after several systems were designed.

In 1973, our company started development of a dictionary system that would provide our "PRIDE" Users with an automated assist to this problem. This product was completed in 1974 and is currently installed in several of our "PRIDE" User locations. We call our dictionary/directory system "PRIDE"-Logik.

As we defined the requirements for "PRIDE"-Logik, we decided that it must include all of the classic properties of a data dictionary/directory system. These are: thorough identification of the physical and logical attributes of each data element; identification of the input documents, output reports and files that contain the data; cross referencing of all these data components; and search techniques that allow the user to identify data elements even though their name or number is unknown.

In addition, we included as requirements the cataloging of all system components identified by the various levels of the "PRIDE" System Structure. This capability permits the cross referencing of systems components with data components. Thus the complete integration of both the systems and data descriptions in one dictionary/directory system.

By developing a system that meets all of these requirements, you can readily see that we mechanically maintain all Information Systems documentation. This approach also allows the system to logically test the threads that link data from its original source through its end use as information.

Since "PRIDE" includes a specific set of activities executed during the design process and these activities force documentation during their execution, the logical checks provide systems quality assurance. These system diagnostics can be performed at several checkpoints during the design process. Each will inform the systems designer of any design problems and allow the elimination of these problems before they become difficult or impossible to correct.

The reaction of our users to "PRIDE"-Logik has been most gratifying. Several large Users with multiple systems and programming groups have commented on the ease of portability of the documentation. Once a system has been designed for one location, it may be proliferated to other locations through copies of the computerized dictionary. Similar portability has also been mentioned by groups having relatively large numbers of personnel working on a single project. In this instance, each project team member can obtain appropriate documentation specifically related to his portion of the project.

Other users have stated that the systems diagnostics greatly assist the process of collecting documentation for systems designed prior to the use of "PRIDE". As the various levels of documentation are prepared, the diagnostics determine whether it is complete or incomplete. Any incomplete areas are corrected and the process continues. All concerned are aware of progress during the execution of the project.

Another aspect often mentioned is the ability to determine the effect of any systems modifications or improvements. "PRIDE"-Logik will identify all systems and data components that may be affected by a change to any other component. This is especially helpful when the contemplated change may cause changes to several systems currently in use.

In summary, an integrated dictionary for systems and data components is not only possible but is a reality. It would not have been possible without a design methodology, system structure concept, data management concept and a complete set of rules for defining systems and data components.

An Information Documentation Language: A Framework for Deriving Information from Data

William M. Taggart, Jr.

Division of Management
School of Business and Organizational Sciences
Florida International University
Miami, Florida 33199

An information documentation language (IDL) builds upon current efforts in the standardization of representations of data elements to recognize the problems of the human decision maker as a user of information processing systems. The standard proposed here focuses on a key phrase from the Federal Information Processing Standards (FIPS) definition for information interchange: "data representing information." Due to imprecision in clearly distinguishing data and information, we have less effective information processing systems from a decision-making viewpoint then many users desire. This paper develops the theme that an information documentation language provides a framework for deriving information from data in order to create more effective information systems for decision makers.

IDL extends the concepts of data element and data item through an information syntax to derive concepts of an information element and an information item. These new concepts provide the framework for an IDL. The precedents for IDL extend back to the early 1960's. These precedents are used as the basis for the synthesis of an IDL framework. After the framework is presented along with a detailed information interchange example illustrating the application of the concepts, both the feasibility and benefits of such a standard are briefly summarized. Efforts of this type will enable us to extend the idea of the management of data elements in information processing to the notion of the management of information in organizations.

Key words: Data definition; data element; data interchange; data item; data standards; decision making; effectiveness; information definition; information documentation language; information element; information interchange; information item; information standards; information syntax; representations of data elements.

1. Introduction

The challenge of efficient and effective information interchange continues to confront the data processing community. We must be increasingly successful in meeting this challenge to help insure the development of information processing systems for a society that is increasingly dependent upon these capabilities for intra- and interorganizational communication. An information documentation language standard as proposed in this paper represents an approach to the effectiveness dimension of this challenge. Efforts of this type will enable us to extend the idea of the management of data elements in information processing to the notion of the management of information in organizations.

1.1. Information Interchange Problem Levels

To place this approach in context, three levels of the information interchange problem may be identified: technical, semantic, and behavioral. [Adapted from 7, p. 4.][1] At the technical level, concern focuses on the accuracy of message transmission. In verbal communication, the problem may be expressed as "Did you receive the same message I sent?" In information interchange, this concern receives the attention of Sections A - Recognition, B - Physical Media, S - Data Communications, and T - Systems Technology of the American National Standards Institute (ANSI) Committee on Computers and Information Processing (X3). [See 10, pp. 99-124 for a summary of X3 activity.] At this first level, the challenge deals with efficiency, i.e. the design of least cost information processing to insure a desired level of communication accuracy.

The second or semantic level of the problem deals with the meaning which the message conveys. At this level, we may phrase the verbal communication question as "Did you understanding what I said?" In the ANSI X3 Committee, Sections J - Language, K - Documentation, and especially L - Representation are concerned in part with the problem of meaning in information interchange. The third level focuses on the behavioral dimension or the effect of the message on the conduct or actions of the message recipient. Here the verbal question might be stated as "Did you behave in the way I wanted you to?" (in response to the message I sent). Both the second and third problem levels deal with the effectiveness of the information interchange, i.e. a consideration of whether the message (output) is understood by the recipient (level two) and achieves the desired result (level three). In this paper, the discussion of an information documentation language (IDL) deals with the second semantic level of the communication problem.

1.2. Information Interchange Definition

To provide a specific point of departure for the consideration of an IDL standard, the following definition of information interchange highlights the issue addressed in this paper:

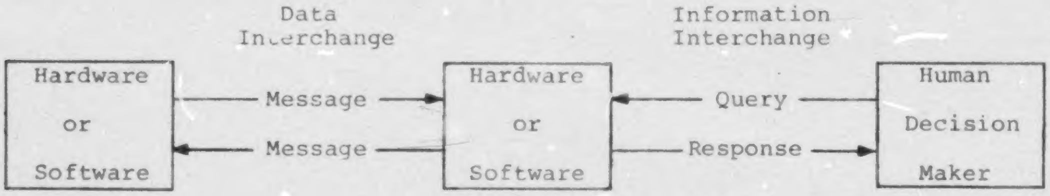
> "The transfer of data representing information between or among two or more points (devices, locations, organizations, or persons) of the same or different (dissimilar) information system or systems." (Emphasis supplied.) [12, p. 8]

This definition suggests the need to understand the way in which data represents (results in) information. The data side of standardization for information interchange represents the responsibility of the ANSI Subcommittee on Representations of Data Elements (X3L8). To complement this data standards thrust, an approach to information standardization should be undertaken to focus on the way data results in information.

---

[1] Figures in brackets indicate the literature references at the end of this paper.

The data and information aspects of standardization can be illustrated in terms of the levels of and definition for information interchange:



As long as messages are being exchanged among two or more nonhuman points in the information system, we are concerned with the accuracy (level one) and meaning (level two) of the data interchange. When we include a human (as a decision maker) in the exchange, an external problem context is introduced where message meaning assumes new dimensions in an information interchange. Here meaning must not only address the common understanding of the message content, i.e. the recipient uses the same data items to represent the data elements as the sender, but in addition provide an information standard for the query and response so that the message will have meaning in the human decision maker's problem context.

1.3. The Need for an Information Standard

At the First National Symposium on the Management of Data Elements in Information Processing, Perry Crawford [2] presented a paper in which he discussed the "inside-out" and "outside-in" approaches to standardization considered in the early work of X3L8. Crawford pointed out that standardization could proceed from the content of fields to the subjects (inside-out) or from the subjects being referenced to the content of fields (outside-in). The "know-what" emphasis of the latter was pointed to as an important complement to the "know-how" emphasis of the former approach which characterizes the current thrust of X3L8 activity. The standard proposed here in the form of an information documentation language seeks to provide a "know-what" thrust to complete our current "know-how" approach.

The need for an information standard may be related to two requirements stated in the Federal Information Processing Standards (FIPS) program:

> "Standards for describing data (information) interchange formats"
>
> and
>
> "Methods . . . for specifying data (information) formats and the data (information) contained therein." [11, p. 5]

In these requirements, the "(information)" interpolation reflects the definition of information interchange quoted above. The need for information standardization implicitly recognized in the FIPS requirements is made explicit when the FIPS information interchange definition phrase "data representing information" is taken into consideration. A standard including a method for the information formats referred to in these requirements is proposed in this paper in the form of an information documentation language. The background survey that follows provides a starting point for an information standard to complement our ongoing data standardization efforts. If pursued such a standard will enable us to deal more comprehensively with the semantic level of the effectiveness problem in information interchange and as suggested in the first paragraph also enable us to begin to develop an approach to the management of information in organizations.

2. The Meaning of Information

In developing information processing systems for decision makers today, we suffer from a lack of precision in the meaning of the term "information." As a result of this imprecision, we do not have a systematic method of dealing with the content of an "information" interchange. We need a standard to more explicitly handle the content of a communication. Viewed from the perspective of the human user, we need to be able to describe the format and to specify the content (as stated in the FIPS requirements) of the message queries and responses of information interchange (as shown on the right side of the diagram on the previous page).

2.1. Data and Information Definitions

As the definition for information interchange suggests, a very close relationship exists between data and information. Here we develop the relation between data and information more clearly as the basis for the specification of an information documentation language standard. One of the more perceptive presentations of the relationship uses the following definitions:

> "Data, the raw material for information, is defined as groups of nonrandom symbols which represent quantities, actions, things, etc."
>
> "Information is data that has been processed into a form that is meaningful to the recipient and is of real or perceived value in current or prospective decisions." (Emphasis supplied.) [3, p. 32]

In the context of a human user of an information processing system, these definitions clearly assert that when we move from the data to information interchange modes of communication we are moving into a decision making environment. An example will highlight potential problems associated with insuring that a user receives information instead of just data.

2.2. Query-Response Example

Consider a situation where a prospective employer (the user) asks a University's student records system "What were Bill Smith's grades for the Information Requirements Analysis course?" In this form the query appears reasonably straightforward. However as is frequently the case in communications, we make assumptions concerning a common frame of reference between the points in the exchange that may not in fact prove valid. Since there is room for misinterpretation, an IDL standard should provide a basis for reducing the ambiguity potential in the message exchange to increase the degree of understanding that results. Let's look at some of the potential ambiguity in this information interchange situation.

First the query is expressed in the natural language of the individual. The flexibility of the natural language structure does not clearly organize the designator and identifier components of the request. There are some implied data elements which identify the person (entity) that the employer wants to know something about. There is another implied data element that designates what it is the employer wants to know about the person. Second the time period for the question does not appear as a part of the request. There is an appropriate period for the activity that the employer wants to know about. In order to explicitly nail down the question being asked, this time dimension will have to be identified.

Third not only must a time period be a part of the request but also the selection of the particular time period may be very crucial. In this situation, the response could be given assuming the academic term when the course was given or what the situation is currently. At the end of the course time period, the individual had an incomplete grade. However since then the incomplete has been made up by finishing the requirements for the course. A response reporting the incomplete may give the impression the individual cannot be expected to complete his commitments. However there may have been very legitimate reasons for the delay. In this case the subsequent completion would be viewed as acceptable performance.

Fourth in the course referred to, there were grades both for individual assignments and for the overall course. The query message seems to imply the overall course grade. In this situation the difference can be very significant. The students contract for an overall grade based on the level of project work they agree to undertake. In this case the student contracted for a "C" due to heavy career commitments at the time the course was taken. However the project work must be completed at an "A" level of performance to be accepted. The individual did complete the agreed upon project work at the acceptable "A" level of performance and therefore successfully completed the course with a "C." Was the individual in question a "C" student or an "A" student? This will remain ambiguous unless the distinction brought out here is clearly understood by the parties to the communication.

A fifth potential basis for ambiguity arises because two Bill Smiths have completed the course. A William R. Smith was enrolled in the winter of 1974 and a William D. Smith, in the winter of 1975. The employer knows which Bill Smith he has in mind, but the student records system will not until this question is cleared up. Finally this course is given at the undergraduate and graduate level with the same title but with a different course number. A different level of performance expectation applies to the two levels of the course. The significance of the grade will in part depend on whether the grade was earned at the undergraduate or graduate level since an "A" in the latter case implies a higher level of professional development in the completion of the project work.

The prospective employer wants to hire a first rate information analyst with the latest skills to work with their systems users to identify information needs. Given the ambiguity potential in this situation, there is a reasonable probability that the response to the employer will be incorrect and that this will not be recognized. As a result, a very serious mistake may be made either in not hiring an individual who should have been hired or in hiring an individual who should not have been. When presented this specifically, the example may appear exaggerated. However each information systems user who carefully reviews their experience can usually provide similar but more subtle and even more significant examples in terms of the decision impact of the consequences of failing to understand an information interchange.

2.3. Deriving Information from Data

In the communication situation just described, much can be accomplished to reduce the ambiguity with the use of standards for the representations of data elements. In fact these standards provide the foundation for the development of the basic concepts of an information documentation language. These concepts and their relationship to data standards may be introduced before going into greater detail on the specific precedents that led to the current formulation of an IDL. Two definitions are basic for the representations of data elements standardization:

> "Data Element - A basic unit of identifiable and definable information. . . ." (Emphasis supplied.)
>
> "Data Item - The expression of a particular fact of a data element. . . ." [12, pp. 7 and 8 respectively]

The three basic concepts of an information documentation language build on these two definitions with the introduction of the following definitions:

> Information Syntax - The essential configuration of data element types which specifies user information structure.
>
> Information Element - A group of data elements that designate and identify potential user information.
>
> Information Item - The expression of a particular fact of an information element.

These definitions provide the framework for deriving information from data. In the next section the specific precedents for the formulation of IDL are summarized. Then in section 4, a synthesis of the above framework is presented in more detail.

3. IDL Precedents

Several lines of development provide the precedents for the information syntax concept of IDL. Its structure follows specifically from five alternative views of data and/or information. This earlier work provides the basis for deriving the data element types of information syntax. The following paragraphs breifly summarize these views in terms of the employer request data elements implied in the above example.

3.1. Information Algebra's Concept of Data

The Language Structure Group (LSG) of the CODASYL Development Committee developed and presented a theory of data processing: Information Algebra [5]. This theory states that more than a single property must be present to represent information. For example, one set of values for the properties 'student number,' 'course,' and 'project grade' might be; '258-54-9119,' 'MAN 621,' and 'A:'[2]

Property	Property	Property
Student Number	Course	Project Grade
Value	Value	Value
258-54-9119	MAN 621	A

3.2. Chapin's Concept of Data

Chapin presented a concept of data which recognizes that any attribute may be the focal property in a data element configuration [1]. A data configuration based on the 'student number,' 'current term,' and 'project grade' attributes would appear as follows:

---

[2] Single quotes are used to identify the data elements and data items used in the examples in this section and the next.

Attribute	Attribute	Attribute
Student Number	Current Term	Project Grade
Value	Value	Value
258-54-9119	Fall, 1975	A

3.3. FIPS Concept of Data

The FIPS standardization program defines a primary data element as one that is the subject of a record and an attribute data element as one that qualifies or quantifies another data element [12, pp. 8 and 7 respectively]. A data configuration based on the 'student number,' 'course,' and 'current term' attributes and the 'project grade' primary would appear as follows:

Attribute	Attribute	Attribute	Primary
Student Number	Course	Current Term	Project Grade
Value	Value	Value	Value
258-54-9119	MAN 621	Fall, 1975	A

3.4. Langefors' Concept of Information

Langefors developed a concept of information which recognizes the essential data element types of information syntax [4, p. 320]. Using the previous configuration, this view of the components of information appears as follows:

Object Identity	Point in Time	Property
		Project Grade
Value	Value	Value
258-54-9119	Fall, 1975	A

3.5. McDonough's Concept of Information

McDonough's concept of information also recognizes the essential data element configuration for information syntax [6, p. 102]. A configuration in his form would appear as follows:

Name	Name	Point in Time or Period of Time	Quality or Quantity
Student Number	Course		
Value	Value	Value	Value
258-54-9119	MAN 621	Fall, 1975	A

3.6. Summary of the Information Syntax Precedents

These examples of five approaches to data and/or information provide the basis for synthesizing the information syntax concept. The contribution of each approach may be summarized as follows (The corresponding information syntax component is shown in parentheses.):

> Information Algebra - More than a single property (attribute, time, and property data elements) should be present in order to identify information.
>
> Chapin - Any attribute (property data element) may be a focal property in a particular data element configuration with the other attributes (attribute and time data elements) serving as identifier properties.
>
> FIPS - Attribute (attribute data element) and primary (property data element) data elements describe the field contents of the records in a file.
>
> Langefors - Three essential components; object identity (attribute data element), point in time (time data element), and property (property data element); are required in the structure of information.
>
> McDonough - More than one name (attribute data element) may be needed to sufficiently identify the set of objects to which a quality or quantity (property data element) refers, and point in time as well as period of time identifiers (time data element) may be necessary.

Bringing these conceptual distinctions together, the information syntax formalizes the structure of a data element configuration to document information interchange from a human decision maker's perspective.

4. An IDL Synthesis

Using the versions of data and/or information summarized in the previous section, the basis for an IDL Standard may be derived. This section presents a synthesis of the preceeding concepts in terms of the information syntax, information element, and information item definitions given at the end of section 2. This synthesis is presented first in terms of the concepts that make up the information syntax and second in terms of the information element and information item concepts which follow from the syntax structure. [This section is an extension of work previously developed by the author in 8, pp. 32-41.]

Three concepts specify data element types for user information structure. One type (property) designates the specific process or state of an entity that interests the information system user. A second type identifies the time (time) dimension of the user's interest. An interest in process or state specifies a period of time or point in time respectively. Finally a third type (attribute) specifies the data element required to identify the specific instance(s) of the entity that the user wants to know about.

The particular data elements for each of these types as determined by the context of a user's particular problem collectively form an information element. This concept name suggests a set of data elements to parallel the data element concept. In turn a set of data items for the data elements form an information item. This concept parallels that of a data item. Since each data element in the information element may assume any one of the data items in its range, a large number of information items may exist for each information element.

4.1. Information Syntax

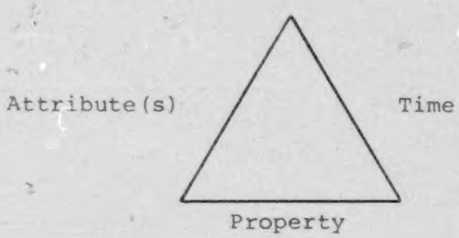
The preceeding section develops a rationale for information syntax as the essential data element configuration which specifies user information structure. Formally information syntax has the following structure expressed as data element types: Attribute(s)/Time/Property. The first two data element types identify the reference entity and the last designates the data element of particular interest to the user concerning the entity. These data element types are formally defined as follows:

Property - A data element which designates a particular process or state of an entity of interest to a user.

Time - A data element which identifies the period of time for a process or point in time for a state of an entity.

Attribute(s) - One or more data elements which identify an entity whose process or state the property data element designates.

These information syntax concepts may be viewed in terms of a triangle which provides stability for the structure of an information query and response for a human decision maker:



If we exclude any one of the sides of the triangle, the information structure of the user's query and response collapses. The property data element provides the base of the triangle by designating the particular interest of the user concerning an entity. To complete a stable structure, we require two types of data element identifiers: time and attribute. The time arm of the triangle represents a special class of identifier which always must be present in an information interchange. Therefore it is singled out explicitly as a part of the syntax. In addition one or more identifiers will be required to pinpoint the particular instances of the entity of interest to the user. The attribute arm of the triangle satisfies this need. If we fail to recognize any one of the three sides of the stable structure in the exchange between user and information system, data will be communicated rather than information. As the definitions for data and information emphasize in section 2, data will not be useful to the human decision maker in his problem context.

The property definition points out that interest in an entity may occur in one of two possible ways. A decision maker may wish to know the condition (state) of an entity at a point in time. Alternatively he may wish to know the change in condition (process) during a period of time. This fundamental difference between state and process property data elements occurs in accounting in the distinction between balance sheet and income statement accounts. The property data elements recorded on balance sheets report the status of entities at a point in time. On the other hand, the property data elements appearing on income statements report on the process occurring for entities over a period of time.

4.2. Query-Response Example Revisited

With this framework for an information documentation language, the ambiguous query-response example presented in section 2 may be systematically approached. First instead of using a natural language structure for the exchange, the information syntax calls for specific types of data elements in a particular arrangement. Second this same syntax insures the inclusion of the time data element that is absent in the original query of the prospective employer to the student records system.

With the requirement to use standard data elements in preparing the exchange, the third problem concerning the term of the course or current term is cleared up. In this case the employer wants to know the grade situation in the 'current term' rather than when the course was completed. Fourth with the recognition of the difference in 'course grade' and 'project grade,' the employer determines the latter to be more appropriate for his decision requirements.

In the fifth possibility for ambiguity, the use of 'student number' (defined as the social security number) clears up the Bill R. Smith versus Bill D. Smith confusion. The 'student number' shows the person (entity) in question to be Bill D. Smith who completed the course in the winter of 1975. Finally with further clarification, we determine that the graduate level version of the 'course' was taken by Bill D. Smith.

The structure enforced by the IDL approach deals unambiguously with the query-response information interchange between the prospective employer and the student records system:

Information Syntax

Attributes / Time / Property

Information Element

Student Number, Course / Current Term / Project Grade

Information Item

258-54-9119 MAN 621 / Fall 1975 / A

Query Response

Using this organizing structure employing the three main concepts of information syntax, element, and item, and using the standards for representations of data elements, an information interchange with a significant degree of potential misunderstanding has been clarified to the extent where a very high probability exists that the exchange will be understood and the parties to the exchange will be confident of this. This contrasts sharply with the situation at the end of the example in section 2.

5. An IDL Standard

5.1. Feasibility of an IDL Standard

In considering the feasibility of an IDL standard, three main phases should be recognized: development, implementation, and maintenance. In development the primary concern asks whether the state of the art is sufficiently progressed to permit the initiation of a standard. If this is the case, then the question of the availability of resources to carry out the development work remains. Given developmental feasibility, will it be possible to implement the standard once developed? This can be viewed in terms of the impact of the standard on suppliers and users and the possibility of legal constraints on implementation.

Once implemented, the work will not be justified if the standard cannot be maintained to reflect changes in the state of the art. This final aspect of feasibility should consider the extent and frequency of change as well as the availability of resources to perform the maintenance activity. In all three phases of feasibility, the cost incurred to achieve the expected benefits must be taken into consideration. At this time the problem has not been explored in enough depth to provide cost estimates.

a. Development Feasibility

The two previous sections summarize the precedents and use them as a basis for the development of an IDL standard. One conclusion is suggested by this material: the conceptual background for an adequate distinction between "data" and "information" now exists. Although this foundation exists, individuals and organizations have not implemented these concepts in the ongoing development of their information processing capabilities. This may be due in part to the absence of a consensus on this distinction. With the development of IDL, changes in practice will more likely occur since individuals and organizations will have an information standard for incorporation in their day-to-day activities in systems development.

Not only must the conceptual foundation be present, but a sufficient number of individuals must be willing to devote their time and the resources of their organizations to the development work. A start at identifying these resources are those individuals who are responsible for bringing the state of the art to where it is today. These individuals represent the core of expertise required to accomplish the development effort. These individuals are most likely to provide consultation and review support for the effort. This implies the need for a small group of users who are individually, and as representatives of their organizations, willing to commit the detail effort necessary to accomplish the development. One purpose in presenting the paper to this group of individuals is to identify the level of interest for this undertaking.

b. Implementation Feasibility

Supplier conformance to an IDL standard arises in the specification of the report outputs of proprietary software packages offered to prospective users. Support for the standard from this sector can be approached in terms of the ability it would provide for presenting an unambiguous statement of the output content of a software package. From a user viewpoint, an IDL standard represents a logical extension of the emerging data standardization efforts to gain control over the specific content of file oriented data processing applications as well as data base management information processing capability.

This standard would provide the user with an unambiuous means for describing the output content of traditional or data base applications as well as a means for specifying new user requirements in a fashion that can be unambiguously translated into specifications for the required input and processing activities. Implementation in the user environment implies policy decisions requiring staff members to utilize the standard in the description of the current content of information processing capability and for specifying new user requirements for processing. From the legal angle this standard would promote a freer market by providing the user with more systematic descriptions of package outputs improving their ability to perform comparative need-have analyses of software acquisitions. This promotes the public interest by enabling more soundly based decisions regarding the purchase of software services from a supplier.

c. Maintenance Feasibility

An IDL standard would have a low degree of volatility. Since the standard would document concepts rather than specific content (e.g. data items) it would not be subject to extensive or frequent revision. The primary maintenance effort would involve infrequent changes to clarify the meaning of the content of the standard as direct user experience raises questions of interpretation. Changes in technology would not affect the specific content of the standard as much as it would the way in which individual users made use of the standard in their information processing environment.

Improved technology would provide improved means for utilizing the standard. Incorporation of these advances would be under the jurisdiction of individual users rather than the responsibility of a standards group. The same individuals who would commit time to the development of the standard would be looked to to provide the resources to perform the infrequent modifications to insure increasing clarity of the content and intent of an IDL standard.

5.2. Benefits of an IDL Standard

The benefits to be realized by the development and implementation of an IDL standard can be viewed from three perspectives: intrinsic, interchange, and educational. Intrinsic benefits are obtained by the user within the context of a specific installation. Next interchange benefits arise from the exchange of information across organizational boundaries either within the same overall organization or between distinct organizational units. Educationally this type standard will provide an accepted basis for developing a clearer appreciation for the student of the relationship between data and information in human decision making.

As with the various aspects of cost in the feasibility section, the current level of problem understanding does not permit the listing of specific economic benefits which will arise from this standards effort. Even when they are better understood, the benefits will be based primarily on intangibles for the information processing community. Benefits will have to be developed in terms of values associated with an improved understanding of the information content of exchanges of information.

a. Intrinsic Benefits

Within an organization the benefits from IDL arise from the ability to build an organizational inventory of the information output currently provided by information processing applications. Developed as a single computerized list of the information elements on all the output reports, this inventory would represent the organization's information menu. [See 9 for the description of an operational version of an information inventory in a University organization.]

Data standards provide for the identification of data ambiguity and duplication in the various files of an organization's information processing applications. By extending standardization to encompass information, the organization will be in a position to identify duplication of information produced by these same applications. The degree of rigor in the structure of the standard implied in this paper makes it possible to computerize the process for identifying this duplication.

b. Interchange Benefits

This area of benefits offers the greatest potential long run payoff for an IDL standard. With the data standardization work as a necessary prerequisite, information standardization through IDL would provide the basis for helping insure that the parties in an information interchange understand what they are communicating with each other. Current representations of data elements standardization efforts are working to insure understanding in data interchange. With this activity well developed, the opportunity exists to begin to deal with standards for information interchange. As this paper suggests, unless the human decision maker's context is encompassed, an interchange represents a data rather than an information communication.

c. Educational Benefits

In the educational environment for individuals preparing for a career in information processing as well as practicing professionals interested in upgrading their skills, a significant need exists to more precisely distinguish the difference between data and information. The IDL standard would provide the logical framework for developing the appreciation of this difference in the educational setting. As this appreciation is developed in the educational context, new entrants as well as practicing professionals will carry this understanding to their work situations. As this evolution occurs, the prospect looms that we may be able to move from an era of data to information processing systems. Professionally we will not really be able to begin to develop information systems on an extensive basis until we appreciate what "information" really implies and provide it as the output of our systems for decision makers. Stated as succiently as possible, we cannot build information processing systems unless we comprehend the meaning of information from the user's point of view.

6. Future Direction

This paper relates the development of an information documentation language (IDL) standard to ongoing standardization efforts in other areas, principally that of data elements and their representations. An appreciation of the significance of the difference between data and information was outlined in the second section. Then the third and fourth sections summarized the precedents of IDL and used this as the basis for synthesizing the foundation for an IDL standard. Finally the last section reviewed the feasibility of the development, implementation, and maintenance of a standard and the intrinsic, interchange, and educational benefits that would result. In outline form, these sections of the paper summarize where we are today with respect to an IDL standard. With this appreciation of the problem of effective information interchange, the question arises of where to go from here.

Your interest is solicited in the initiation of an ANSI standards development project with the purpose of pursuing the material in this paper in sufficient detail to determine the merit of an IDL standard for the information processing community. The goal of the effort will be to develop the phrase "data representing information" found in the FIPS definition of information interchange to apply in the problem context of the human decision maker. This appears as an increasingly significant need for the information processing community. The scope of the effort will be to develop an IDL standard closely integrating this with the ongoing work of related ANSI X3 activities. The program of work will involve the analysis and development in greater detail of the feasibility and benefits of an IDL standard for information interchange. Efforts of this type will gradually enable us to extend the idea of the management of data elements in information processing to the notion of the management of information in organizations.

7. References

[1] Chapin, Ned. "A Deeper Look at Data." Proceedings of the 23rd National Conference of the ACM. 1968.

[2] Crawford, Jr., Perry. "On the Connections Between Data and Things in the Real World." First National Symposium on the Management of Data Elements in Information Processing COM 74-10700. Washington, D. C.: U.S. Department of Commerce, April 1974.

[3] Davis, Gordon B. Management Information Systems. New York: McGraw-Hill Book Company, 1974.

[4] Langefors, Borje. Theoretical Analysis of Information Systems. 4th ed. Philadelphia, Penna.: Auerbach Publishing, Inc., 1973.

[5] Language Structure Group of the CODASYL Development Committee. "An Information Algebra." Communications of the ACM 5 (April 1962): 190-204.

[6] McDonough, Adrian M. and Garrett, Leonard L. Management Systems: Working Concepts and Practices. Homewood, Illinois: Richard D. Irwin, Inc., 1965.

[7] Shannon, C. E. and Weaver, W. The Mathematical Theory of Communication. Urbana, Illinois: The University of Illinois Press, 1949.

[8] Taggart, Jr., William M. "A Syntactical Approach to Management Information Requirements." PhD Dissertation, University of Pennsylvania, 1971.

[9] Taggart, Jr., William M. "Developing an Organization's Information Inventory." Management Informatics 3 (1974): 283-92.

[10] U.S. Department of Commerce. National Bureau of Standards. Federal Information Processing Standards Index. FIPS PUB 12-2. Washington, D. C.: Government Printing Office, 1974.

[11] U.S. Department of Commerce. National Bureau of Standards. Objectives and Requirements of the Federal Information Processing Standards Program. FIPS PUB 23. Washington, D. C.: Government Printing Office, 1972.

[12] U.S. Department of Commerce. National Bureau of Standards. Standardization of Data Elements and Representations. FIPS PUB 28. Washington, D. C.: Government Printing Office, 1973.

8. Addendum

Question Give an example of data having no informational content.

Answer As Harry White pointed out in his presentation to the Symposium, information is in the eye of the beholder. This observation is a basic premise of an Information Documentation Language (IDL). For instance consider the following situation. A human decision maker may consider (in his windowless office) whether it is raining outside. A degree of uncertainty exists in the mind of the decision maker with respect to this state of nature. An associate passes by the door of the individual's office and casually mentions that it is now raining outside.

This interchange reduces the individual's uncertainty but does not represent information since it is not related to the decision maker's present or future decisions in carrying out his organizational responsibilities. That is the data items received [now (time data item) raining (property data item) outside (attribute data item) in terms of the IDL syntax] do not match up with a decision problem or organizational responsibility of the user. In contrast if this group of data items did relate to the discharge of a decision responsibility, then what was merely a data interchange becomes an information interchange.

This example truly illustrates that information is in the eye of the beholder. However the IDL syntax provides an organizing framework for insuring that when the beholder's eye perceives a problem relevant string of data items the necessary components will be present [property, time, and attribute(s)] to derive the needed information item.

Question Isn't "time" logically an "attribute"? Could you explain why you treat it seperately? Is it because it is so often overlooked?

Answer Both of your observations are correct. In effect the time component of the IDL syntax is a special case of the attribute component. Since it is an attribute that must always be present and is in fact so frequently overlooked, special treatment is given to this "attribute" by singling out "time" as a distinct component of the IDL syntax.

Question I am not sure whether the speaker was promoting the development of a new software or informing me of a model in existance. If there is a model or pilot being developed, could you tell us where.

Answer A pilot has been developed using IDL. This project is described in "Developing an Organization's Information Inventory" [9]. This reference also presents IDL from another viewpoint as it especially relates to the information inventory problem.

This paper proposes to extend the IDL concept further as the basis for initiating the development of information standards to complement the current efforts underway in the development of data standards. It is my contention that we must eventually take the data processing user into account as a human decision maker if we are to ever successfully come to grips with the effectiveness dimension of information interchange in organizations.

Question Please comment on the relation of IDL to data dictionaries; in particular as to data definition as related to usage.

Answer IDL depends on the concurrent development of data standards. In the application of IDL, the data elements and data items come from a data dictionary which reflects the data use identifiers appropriate to the decision makers in the organization. The data elements and data items used in the example in section 4.2 were taken from a data dictionary developed to support the implementation of the information inventory mentioned in the answer to the previous question. The data elements used must have definitional meaning for the decision maker in question.

Question In your opinion, what or how much background does a "manager" need to have to effectively work with IDL?

Answer It depends on the use made of the IDL concept. If the objective is to develop an organizational information inventory, then the systems analyst can work with the advice of the appropriate managers in identifying the information content of reports without the managers having to understand the IDL concept in detail.

On the other hand if the purpose is to use IDL in the context of identifying a manager's new or modified information requirements, then the process will be facilitated if the manager has at least a basic understanding of the IDL concept. My experience with MBA evening program students who are managers or managers to be indicates they can grasp the essence of the concept easily and in some cases more readily than systems personnel. The managers' more natural orientation is to think in terms of information. Unfortunately for the effectiveness of the data processing profession, the systems professional too often thinks *only* in terms of data.

International Standards for Data Transmission

V. N. Vaughan, Jr. - Chairman CCITT Study Group Sp A

A.T.& T. Co.,
195 Broadway,
New York, N. Y. 10007, U.S.A.

Commencing its work in 1960, CCITT Special Study Group A on data transmission has developed approximately 35 standards for data communication over the public telecommunications networks. These are widely observed in all of the developed nations of the world and have provided a foundation for the growth of teleprocessing systems. These standards include interfaces between data processing and data communications equipment, transmission bit rates, transmission codes, modems, transmission error control, acoustic coupling, the use of digital facilities, switched network and leased line applications, etc. The role and responsibilities of CCITT and the relationship of its work to other standards organizations are described. The accomplishments of Sp A are summarized and some insights are provided on current and future activities in this field.

Key words: CCITT, code, data transmission, interface, international alphabet No. 5, ITU recommendations, standards, switched network, telecommunications.

1. Introduction

This paper will describe principally the work of CCITT (an abbreviation for the French words meaning Telegraph and Telephone International Consultative Committee) in the field of data communications. To set this work in perspective, it is first necessary to know something of the normal role and responsibilities of CCITT and how it relates to other domestic and international standards activities.

The CCITT is an arm of the ITU (International Telecommunications Union) which is a treaty-level organization established over 100 years ago. It is related to the United Nations and recognized by the UN as being responsible for all international telecommunications matters. ITU conferences are held about every six years and are attended by large official US delegations headed by an ambassador-level representative of the Department of State. Included are representatives of the FCC and the major US domestic and international telecommunications carriers. CCITT Plenary meetings are held, usually at the ITU headquarters in Geneva, Switzerland, about every three years and ratifies the work of its study groups.

CCITT "standards" are not correctly called standards but are termed recommendations. This terminology is essentially a euphemism because, in most countries of the world, these recommendations are rigidly observed - more so than most other standards. In CCITT parlance "Recommendation X.27: Electrical Characteristics for Balanced Double - current Interchange Circuits. . ." is equivalent to or synonomous with "Standard X.27: Electrical Characteristics. . .".

The substantive work on "standards" in CCITT is done by its 16 numbered study groups (I through XVI, which cover tariff principles, telephone switching and signalling, telegraph switching and signalling, etc.) and three Special study groups (Sp A, C and D). Special Study Group A (Sp A) is responsible for data transmission. It was the first CCITT study group to be designated "Special". The reason for such a designation was that the CCITT plenary which authorized Sp A in 1960 recognized that its work would be of a broad character and would overlap the work of several existing "numbered" study groups working on both telephone and telegraph matters. This has proven to be the case and SP A has cooperated with and drawen support from a number of the telegraph, telex and telephone study groups.

The first activity of Sp A was to elaborate on its work program to cover in detail subjects such as network characteristics, interfaces, transmission speeds, codes, modems, error control, acoustic coupling, automatic calling and answering, use of group bandwidth circuits, the use of digital facilities, switched network and leased line applications, etc. In proposing answers to these study questions, the members of Sp A have submitted documents totaling many thousands of pages during each plenary period. Discussing and debating the merits of the various proposals and arriving at a consensus, hopefully unanimous agreement, has consumed many hours of the delegates' time both in the meetings and in private discussions where, it can often be truthfully said, the real understanding and negotiating work takes place.

The meetings of Sp A and its several WP's (Working Parties) have all been held, with only one exception, in Geneva at intervals of six to fifteen months. Reflecting the intense and widespread interest in data transmission, all have been relatively large meetings. With an attendance often exceeding 200, typically there have been representatives from the PTT's (Postal Telephone and Telegraph administrations) and RPOA's (Recognized Private Operating Agencies or Telecommunications Carriers) of all the larger developed nations and many of the small highly industrialized countries. These experts have been assisted by about an equal number of representatives of computer, modem and data terminal manufacturers plus other international organizations.

Interestingly and closely related to Sp A, the most recent of the "special" study groups to be established, in 1968 is Special Study Group D. Sp D is charged with setting standards for digital facilities. Also closely related is SG VII, established by the 1972 plenary in Geneva, an outgrowth of a Joint Working Party of SG Sp A and SG X (Telex Switching). SG VII was chartered to establish standards for new public data networks, based primarily on the use of digital line facilities. Naturally there is some overlap and a high degree of coordination required between the work of Sp A and SG VII.

2. Other International Activities on Data Transmission

In the same year that Sp A began its work, a very closely related activity began in the International Standards Organization (ISO). Technical Committee No. 97 (TC 97) was established to cover computers and information processing with Sub-committee 6 (SC 6) charged with data transmission. While a superficial view might suggest a complete overlap and duplication of the work of ISO TC97/SC6 and CCITT SG Sp A, in fact there has been a relatively small overlap and virtually no duplication of effort in the two organizations. This is due largely to the cooperative spirit and efforts of the leaders and workers in both organizations who have seen their respective charter responsibilities and spheres of competence as being complementary rather than competitive.

CCITT Recommendation A20, first approved in 1964, revised in 1968, again in 1972 and likely to be further refined in 1976, lays down the basis for this cooperation and division of responsibilities between ISO and CCITT. Briefly, ISO is responsible for the standards relating to customer-provided computers and data terminals which connect to telecommunications while CCITT is responsible for the standards for the telecommunications network interfaces and for the telecommunications networks including transmission channels, switching, signalling, etc. For example, while CCITT has adopted a transmission code for information interchange, it has done no work on coding of country names for use in data processing. This has been done by ISO TC97/SC14 (supported by ANSI X3L8). Needless to say, as technology and the state-of-the-art advances in both data processing and data communications, expanding and refining the division of responsibilities between these two organizations is a continuing process.

Sp A has also been aided in its work by the cooperation and active participation of a number of other international organizations including ECMA, IATA, ICAO, IPTC, WMO and UIC. With these organizations, the division of responsibilities has been fairly obvious.

3. Relationship of U.S. Standards Work and CCITT

In the field of data communications, there is interaction and a close relationship of the work done in the USA by ANSI and EIA to the work of CCITT. Paridoxically, however, neither ANSI nor EIA has any formal relationship with CCITT nor would either one be eligible for membership. There are essentially three classes of CCITT membership: (1) governments (often represented by the PTT) and RPOA's, (2) scientific or industrial organizations and (3) international organizations. In the first category are the U.S. Government, AT&T Co., WU, WUI, etc. In the second category are IBM, Honeywell, GE Co., etc. In the third category are ISO, ECMA, etc. There are about 30 U.S. members of CCITT. But national standards bodies are not members of, or participants in, the work of CCITT.

To fill what would otherwise be a gap in official communications between the work of ANSI (viz. X-3S3), EIA (viz. TR30.1) and CCITT, the U.S. Department of State has established an Industry Advisory Group with several study groups including SG No. 5 which is responsible for data transmission. Meetings are held in Washington, as required - usually several times a year, and are open to the public. The Chairman is Mr. T. deHass, Institute for Telecommunications Studies, U.S. Department of Commerce, Boulder, Colorado 80302. Attendance usually includes representatives of all the U.S. domestic and international telecommunications carriers, the leading business machine companies, modem and other data communications equipment manufacturers.

This group develops the USA position on any CCITT activity on which there is no other ongoing U.S. standards activity, viz. modems. It also approves and officially forwards to the CCITT any relevant positions that have been developed in ANSI or EIA, viz. the work on the interface with new public data networks which is in X3S37. Additionally the work of ANSI often also officially reaches CCITT via ISO which is an active liaison member of CCITT.

4. Network Characteristics

CCITT SG Sp A began its work with a study of the feasibility of data transmission over the general switched telephone network and leased voice-grade lines. Today, looking back, it is hard to believe how relatively little was known in 1960 about the end-to-end performance of the telephone network. Many respected experts were of the strong opinion that it would be impossible to achieve satisfactory data transmission over it. Various types of modulation and modems (modulator/demodulator) were studied and experimentally tested at various operating speeds (bit rates) to determine the overall performance in terms of bit error rates. Supporting this, there have been extensive measurements made of the network characteristics in terms of analogue parameters such as attenuation, delay distortion, noise, impulse noise, phase jitter, non-linearity, etc. This data has provided the foundation for the work of Sp A, especially in its work on modems. Most participating countries have rather completely reported on the performance of their national networks, some having updated their reports one or more times. There have also been reports on the end-to-end performance of a good statistically sampled number of international connections. Reference: Supplements in Red Book Vol. VII, White Book Vols. IV and VII, Green Book Vols. IV and VII. These are publications of the ITU. A complete list of ITU publications with prices will be sent free of charge on request to the General Secretariat, International Telecommunications Union, Place des Nations, Ch-1211 Geneva 20, Switzerland.

In the early years of data transmission development, the simpler types of modulation (viz. AM and FM) were used. Therefore, the data transmission experts were interested in only a few parameters (such as impulse noise) which are of relatively small concern for voice transmission. It was only later as more sophisticated types of modulation came into consideration (4-phase, 8-phase, AMVSB, SSB, QAM, etc.) that interest centered on some previously unknown or ignored parameters such as phase hits, second and third order cross-modulation products, etc. Not only has it been necessary to define new parameters but also to develop methods of measuring them. This has sometimes amounted to the standardization of a new test set. For example, an impulse noise measuring set and a phase jitter measuring set have been defined. While the formal approval and issuance of all recommendations on testing and maintenance are the responsibility of SG IV, (maintenance), it has been Sp A which has had the motivation and provided most of the technical expertise needed to accomplish the needed result in the field of data transmission.

Further, uniform methods of expressing performance have been defined and continue to be refined. For example, initially the error rate was expressed only in terms of the long-term average bit error rate, viz 1 x 10^{-5}. Subsequently, the use of sequence error rate, often called block error rate, has come into use, viz 2% blocks in error for blocks of 511 bits in length at 1200 bps. Consideration is now being given to a definition of error performance in terms of % error-free seconds for given speed, viz 99% error-free seconds at 4800 bps.

In the case of the switched telephone network, because end-to-end connection characteristics vary depending on locations and route selection, it is necessary to express performance statistically. For example, 90% of the connections over the Federal Republic of Germany network will have a block error rate of 10^{-2} for a 512 bit block length at 1200 bps. (Ref. Green Book, Vol. VII, p. 306). A large amount of such information is included in the Red, White and Green books.

In the case of leased voice grade lines, the end-to-end characteristics have been described for data-grade channels in terms of minimum and maximum limits for all of the relevant analogue parameters. Initially these were established in 1968 and known as M-88, revised in 1972 and known as M-102. Currently, because most modern modems use automatic adaptive self-equalizers, study is focused on a new definition of attenuation and delay distortion which will not be held to such tight values (because meeting tight limits often results in many "wiggles" in the circuit characteristic) but looking towards a smooth and broad bandwidth characteristic. An in the case of test sets, the official responsibility for issuance of these recommendations belongs to SG IV but Sp A is instrumental in their origination and preparation.

The performance of the switched networks and leased lines has substantially improved during recent years, both for data transmission and for ordinary telephone usage. This is due in part to the focusing of attention on performance that has come as a result of the use of the network for data transmission. The troublesome conditions that have been corrected were not caused by data transmission but were found because of data transmission. It has been said that data transmission is to the network like the thermometer in the patient's mouth.

5. Interfaces, Speed and Code Standards

As a very practical matter, in order for the manufacturers of data terminals and computers - and their customers - to be able to connect to and use the telecommunications networks, the interfaces between the terminals and networks must be well defined. The work on interfaces began in the CCITT Working Party No. 43 on Data Transmission which preceded the establishment of Sp A and has continued actively in Sp A to this day. During the current 1972-1976 Plenary Period a new (Provisional) Recommendation has already been approved for interface electrical characteristics which are compatible with integrated circuits.

Work on this network interface, between the "Data Communications Equipment" (DCE) and the "Data Terminal Equipment" (DTE), was a significant extension of the previous scope of CCITT studies. The primary CCITT efforts have been to define signals at the international frontier to insure compatible interworking while leaving complete freedom of design for the national networks and local stations. It was clear that the emergence of customer provided computers and other data terminals required the standardization of the DCE/DTE interface in order to achieve interoperability.

This work on both functional and electrical interface specifications has always had and continues to have the close cooperation and valuable assistance of ISO TC97/SC6. Since this interface is essentially the line of demarcation of responsibility between customer-provided equipment and the telecommunications network, also between ISO and CCITT, the importance of this study and its results (Recommendations V.24, V.25, V.28, V.35, X.26 and X.27) cannot be over-emphasized. It has permitted the same type of data equipment or computer port to connect to and operate satisfactorily with the network in all of the participating countries. The computer or terminal is able to manually or automatically originate and answer a data call and communicate via the standardized modems.

Because of the wide variety of local, national and internationsl transmission and signalling systems, the successful development and use of this interface was no simple matter. Some considerations which guided the work were: (1) what customer service features are needed, (2) maximum freedom of design should be possible on either side of the interface without affecting the other side, (3) clear definition and measurable performance characteristics are required at the interface, (4) it should be as simple (economic) as practicable and (5) each party (viz. ISO and CCITT, or the customer and PTT) will have some interest in, but not responsibility for the design and performance on the other side of the interface. Incidently, these same considerations are guiding SG VII in the establishment of the X.21 interface for new switched public data networks.

Obviously another of the earliest needed essential standards is for data signalling rates (speeds) in terms of "bits per second" for synchronous operation. These standards were set with a number of considerations in mind including the network performance capability, modems, and economics. A limited set of speeds, which are multiples of 600 bps, up to 9600 bps have been set (V.22). Because the future use of higher speeds is expected to be predominantly over digital facilities, speeds above 9600 bps are based on a limited set of the multiples of 8 kbps. Only 48 kbps has yet been set.

To permit the automatic setting up, answering and clearing down of data calls over a switched network obviously requires use at the DTE/DCE interface of a standardized code, or "alphabet" as it is called in CCITT. At the time ISO was developing a new code for information interchange (an equivalent of the ASCII code in the USA), SP A recognized that the then most modern CCITT code (a 5-bit code, I.A. No. 2) would be inadequate for future interworking and teleprinter terminals. Accordingly, it joined with the already well advanced ISO work and the two organizations produced a joint standard which meets the needs of both the communications and data processing communities. This is known in CCITT as I.A. No. 5 (International Alphabet No. 5).

6. Modems

At the very heart of Sp A's work is the subject of modems. The modem interprets, ie, interfaces, the data terminal to the network and vice versa. It determines, one might also say controls, the end-to-end data transmission capability which the data terminals or computers can use. Furthermore, at its interface with the data terminal, it is possible to give a much more succinct and simple description of the end-to-end data transmission performance in terms of bit rates and error characteristics than would be possible in terms of analogue parameters.

The stranger to this work might ask: Why has Sp A standardized so many modems? Why is it not possible to design a single general purpose modem for all applications? It would be about as practicable as a one-design truck for all applications. There is too wide a range of traffic volumes, message lengths, speed-of-response needs, economics and customer service feature requirements to make a one-design approach practicable. All of these factors must be considered in the development of modem standards. What is needed is a family of modems, to be kept as small as good family planning will permit, which can meet all of the needs with a near-optimal balance between economics and performance.

Additional factors to be considered in modem design include echo suppressors and signalling systems used in the switched telephone network. Consideration had to be given, not only to the international signalling systems, but also to the many different national signalling systems in existence. Without due consideration to these factors, portions of the data message might be blocked by echo suppressors or the connection might be automatically broken down due to interference with a signalling system.

Especially with the high-performance modems, the greatest difficulty in reaching agreement on modem standards is often in choosing between two seemingly equal designs where each has a different superiority (viz. tolerance to phase jitter or bandwidth required) and a compromise is not technically possible which combines all the good attributes of both designs. Further, some designs which have had the highest attractiveness when initially studied from the standpoint of theoretical considerations have subsequently been found least desirable when state-of-the-art implementations (another time variable) are considered.

Another problem is that of how to write a recommendation in such a manner that it adequately describes the "signal at the international frontier" - so that different manufacturers in different countries can build compatible modems - and to avoid the recommendation being a hardware specification, something which is not within the scope of CCITT jurisdiction. There are also a few possible pitfalls in the patent area which must be carefully observed.

The foregoing tends to make the establishment of modem recommendations seem to be nearly impossible. Sometimes when conflicts of commercial and national self-interests are added to the mix, it seems truly impossible. However, that it is possible is attested to be the substantial number of modem recommendations which have been approved and by the widespread, successful and rapidly growing use of these recommendations.

The successful interpretation and translation of modem recommendations into compatible hardware designs has not always been without problems and it has sometimes been necessary to modify or expand an approved recommendation in order to relieve some compatibility problem. But even the best of living standards are usually susceptible to fine-tuning as the state-of-the-art advances and as experience is gained in their application.

7. Modems for the Switched Telephone Network

The first modem approved by Sp A was fairly simple FM modem (V.23) for operation at speeds up to 600/1200 bauds assynchronously or 600/1200 bps synchronously. It is a one-way-at-a-time modem but reversible with only a few milliseconds required for reversing under control of the data terminal. There is optionally a separate reverse channel, which can be operated at speeds up to 75 bauds, intended to be used for error control. 1200 bps is considered the nominal operating speed with 600 bps a fall-back for those connections or locations which are incapable of operation at the higher speed.

Since the initial adoption of V.23 in 1964, this recommendation has been supplemented to include automatic calling and answering, as well as an optional clock and echo suppressor control. The ISO, together with many computer and terminal manufacturers have developed extensive and highly satisfactory protocols for the control and use of the telecommunications network with these modems and their flexible set of capabilities. Indeed, operation with the more recent and sophisticated higher speed modems is largely based on the protocols developed for use with V.23. In retrospect, computer/terminal protocols and software have seemed more durable and fixed than modem hardware.

During the same 1960-1964 plenary period, another very fundamental recommendation developed was for a FDX (full-duplex) modem (V.21) capable of simultaneous two-way transmission at speeds up to 200 bauds. The two directions of transmission are made independent of each other by means of frequency division. As with V.23, this recommendation has been refined, supplemented and improved during each plenary period since its inception and the process is continuing.

The extensive development and use of low-cost teleprinter terminals, line protocols and computer software for use with the V.21 modems makes it desirable to increase their maximum speed capability. This will probably be increased to 300 bauds during the current plenary period (1972-1976). This appears to require a relatively small supplement to V.21 and interworking between the existing modems and new modems should be possible at 200 bauds. 300 bauds would be the limit when two new modems are interworking. Ultimately, the older modems will probably be phased out.

A further extension of the speed capability for full duplex operation over the general switched telephone network is also currently under study. By utilizing a more sophisticated modulation technique (viz. 8-phase modulation with automatic equalizers, encoders and/or scramblers, with frequency division between the two directions of transmission), it appears possible to achieve the same operational capability as with V.21 at speeds up to 1200 baude or 1200 bps.

In 1972 approval was given to 2400 bps operation over the switched network using a 4-phase modem (v.26 bis). The f[illegible]-back speed is 1200 bps. Except for the type of modulation, speed and higher cost, this modem is operationally comparable (end-to-end, or interface at one end to interface at the other end) to V.23.

During the current plenary period, intensive study abtivity is directed at 4800 bps operation over the switched network with a new modem operationally comparable to V.23. Clearly, automatic adaptive equalization and scramblers will be required to uniformly distribute the signal energy over the bandwidth and avoid sensitivity to certain combinations. AM VSB, QAM and 8-phase modems are strong candidates and are being compared. With a crucial meeting of Sp A at which this study will be a principle subject scheduled for Geneva in March 1976, it would be indelicate (to put it mildly) for the chairman to predict the outcome at the time of this writing.

8. Modems for Leased Voice-Bandwidth Lines

The main and most important item that the manufacturer's representatives to Sp A would have us remember about operation over leased lines is that the customer should always have complete freedom of choice to use any type of modem he chooses, regardless of whether or not it conforms to a standard, just so long as its total power level output does not exceed the constraints of V.2 (-15 dBm0 for each direction of transmission, with a few detailed caveats). This viewpoint has generally been accepted by Sp A and by most of the PTT's and RPOA's. However, there has also been acceptance of the need to have available standardized modems which represent the state-of-the-art, optimally best choice for most applications over two-point and multi-point leased lines.

While most manufacturers and several international organizations have championed customer freedom of choice, they have also been among the most active participants in the work and have demonstrated a highly conciliatory spirit in the traditional CCITT manner of reaching compromises which are often necessary to achieve a standard.

The principal difference between the leased line and switched network operation is that leased lines are assumed to be four-wire circuits end-to-end permitting simultaneous and independent usage of the full bandwidth in both directions. Another difference is that there are no constraints imposed by telephone signalling systems or echo suppressors. However, there has been a strong tendency in Sp A to prepare recommendations for leased line operation with a view towards the same basic modems being a subsequent candidate for operation over the switched network. Perhaps this latter point has been one of several factors which has stimulated many modem manufacturer's interest in the work of Sp A.

V.23 modems, operating at 1200 bps FDX, have been a standard for leased lines since 1964. In 1968 a 4-phase modem, V.26 with option A and option B, was adapted for 2400 bps usage in an operational mode comparable to V.23. This led to the adoption in 1972 of V.26 bis, with option B preferred, for switched network operation. The existence of options A and B in V.26 is illustrative of one of the difficulties sometimes encountered in reaching complete agreement. The encoding of the dibits 00, 01, etc., corresponds to a phase change of 0° , + 90°, etc., in option A, but to +45°, +135°, etc., in option B. Each option has certain small but important transmission advantages over the other, but the two are not compatible. The 1976 plenary is expected either to confirm option B as the sole standard or adopt a replacing higher performance modem for operation at 2400 bps over both leased lines and the switched network.

The V.26 modem is also a good illustration of the continuing nature of standards work on problems that were unforeseen when it was orginally adopted. Almost six years after V.26 was adopted, with perhaps 50,000 modems built by a dozen manufacturers around the world in service, severe compatibility problems arose between those of one manufacturer and most of the others. It required considerable testing, analysis and study to determine that all the modems complied fully with the recommendation but that the recommendation was deficient in not covering, in some way, limits on the distribution of delay distortion between the filters in the transmitter and in the receiver. This deficiency is currently being corrected by Sp A.

In 1972 an 8-phase modem with scrambler (V.27) was adopted for operation at 4800 bps over leased lines. Although this recommendation is not considered as quite completed, because it does not include an optional automatic equalizer which is recognized as essential for some applications (viz. polling on multi-point lines), it is being widely use quite successfully. Work is continuing during the current plenary period to further refine this recommendation.

9. Data Transmission over 48 KHZ Group and 240 KHZ Supergroup Circuits

In 1968 in Mar del Plata, the CCITT plenary approved Sp A's recommendation V.35 covering the transmission of 48 kbps, plus a V.F. control channel, customer premises to customer premises. Subsequently, both the required analogue characteristics of the 48 kHz channel and the definition of the modulation process have been refined. Work is continuing to permit operation at higher speeds (viz. 64 kbps) and provide new interfaces more appropriate to forthcoming applications. These new applications are principally as off-net bearer channels for off-net extensions of the new digital networks and higher capacity common channel signalling systems to be used for both telephone and data services.

The guiding fundamentals in planning for data transmission over group and supergroup bandwidths are not very different from those for high performance modems to operate over voice bandwidths. In general, one wants the highest possible bit rate consistent with complexity (ie, economics) and an error rate which is low enough to achieve the highest net throughout of error free information bits (after the redundancy for error control has been deleted). There are, however, two additional considerations not of significance in voice band operation. The amount of spectral energy which is concentrated in any 4 kHz portion of the channel should not exceed one-twelfth of the total power. This is to avoid interference into voice circuits which share the same supergroup facilities. Also, when using 48 kHz circuits, one must be even more careful than with voice bandwidths to minimize spurious out-of-hand signals.

The possibility of operating two 64 kbps services or one 128 kbps service over a 64 kHz channel is also being studied. Study of the use of 240 kHz circuits is mostly awaiting economic justification and needs for data transmission at speeds above those obtainable over group bandwidths.

10. EKG, Acoustic Coupling, Error Control, Etc.

The full scope of the work of Sp A could only be covered by reference to the CCITT Red Book (1964) Vol. VII, White Book (1968) Vol. III, Green Book (1972) Vol. III, plus the nearly one-foot stack of white, delayed and temporary documents issued by the secretariat to the members of the study group during the current plenary period. A paper as short as this one could not possibly do justice to the hard and effective work of the many participants, the organization of the work, or the many subjects studied. However, there are a few subjects of study, not previously touched upon here, which should at least be mentioned.

In keeping with Sp A's continuing desire to provide for all reasonable customer service needs, in 1972 recommendation V.15 was adapted to provide for the acoustic coupling of data terminals to the switched telephone network. Study has been initiated on the methods of transmitting electrocardiogram signals, either as analogue signals or as digitally encoded signals.

Although recommendation V.41 covering a code-independent error control system was adopted in 1968, most members of Sp A have felt that the great majority of error control systems used would be internal to the data terminal or computer, ie, on the customer's side of the interface. Therefore, in the study of error control, emphasis has been put on the provision of transmission and modem facilities which permit efficient use of an end-to-end error control system. However, possible new error control systems are under continuing study in Sp A, particularly in relation to the use of satellite circuits with their much longer propogation delays than terrestrial circuits.

Other recommendations adopted concern maintenance methods and procedures for making comparative tests on modems. A parallel data transmission system (V.30), with all the bits comprising a character being transmitted in parallel (ie, at the same time) by means of frequency division, rather than serially, has been adopted. Study is continuing on this method of transmission with emphasis on the use of the 12-button telephone "dial" as an input device for data transmission.

11. Conclusion

The past 15 years have witnessed the emergence of data transmission from the laboratories and from a few specialized government applications into a big and still rapidly growing business. The use of existing telecommunications networks has facilitated this growth without requiring the high initial getting-started costs inherent in developing and building a new network especially designed for data transmission. That the data transmission business has now grown to such proportions that new data networks are coming into being, is partially, at least, a result of the work of Sp A. The availability of the existing networks plus the driving force of computer automation have also been big factors in the growth of data transmission.

In the future, even after the new data networks become available, because of the ubiquitousness and flexibility of the telephone network, its attractiveness for many applications of data transmission will remain indefinitely. Its attractiveness is enhanced by the decreasing costs and increasing performance capabilities of modems.

One might think that, having worked for 15 years on data transmission standards and having transferred work on the use of digital facilities to SG VII in 1968, Sp A would find itself without much to do today. But quite so the contrary, the current study on how to standardize the transmission of 9,600 bps over leased lines is as exciting (and as difficult) as any of the past work of Sp A. The same can be said for 1200 bps FDX and 4800 bps HDX, over the switched network. There are even suggestions of using adaptive echo cancellation techniques and thereby achieving 9,600 bps FDX operation over the switched network. The end is not clearly in sight. While it would be interesting to speculate about progress to be made during the current and next plenary periods, to do so could be counterproductive.

In its study, Sp A has drawn heavily on the thoery and technology used in telephony, telephone signalling, and telephone transmission. Many experts from these fields have contributed to the work. The computer and data terminal manufacturers have made valuable contributions, not only to technical solutions but more importantly to an understanding of customer needs and the market for features and functions. Undoubtedly it is the diversity of participation as well as the depth of participation which SP A has experienced that is necessary to achieve satisfactory results in international data transmission standardization.

Sp A has worked with a rapidly developing state-of-the-art. It has faced and will continue to face the age-old dilemma known to all who work on standards. To set a standard too early can stifle further technological development but if the effort is made too late it is almost certain to fail, because a variety of "de facto" standards will be well entrenched. The problem is one of timing. That Sp A has largely overcome this dilemma is probably due to its collective sense of pragmatism in preference to perfection.

AN INFORMATION MANAGEMENT VIEW OF DATA MANAGEMENT

Marvin G. Wallis

Reports and Project Management Division
Information Systems Office
National Aeronautics and Space Administration
Washington, D. C. 20546

Data management has become increasingly difficult in today's world with the increased usage of computers and the ever-increasing demands of users for data of all types. We have often, in our haste to meet these demands, failed to reduce our requirements for data to that which is needed and have acquiesced in the production of data which "might be nice to have". Unable to control, or prohibited from controlling, the flow of information to meet our real needs, we have reached for a solution and found the smallest piece of the problem. We have arranged for various types of data element management through standards, categorization, definition, and control. This does not become, in itself, an evil but simply avoids meeting the real problem which is the flow of information on a timely, useful, and needed basis. This is an attempt, based on our own experiences, to again bring the larger problem into our scope, to allow us to reconsider the matter, and to present a possible approach to providing a solution to meet the requirements of a vital, and viable, information system.

Key words: Data element management; data management; information flow; information management; information system.

1. INTRODUCTION

Increased use of data banks and automated data processing has led to some consideration of data element standardization. Government/industry cooperation has gradually broadened the acceptance of such standards and of hardware and software standards as well.

Often, data management today seems to place its emphasis upon the data element. Today we see "data administrators" and "data element administrators". We hope such increasing specialization will not result in job titles or descriptions as "bit" or "byte" administrators, for example.

While we realize that there is a need for standardization of data elements and for certain hardware and software standards, we often wonder if we have concentrated our efforts on the small bits and pieces, considering them "manageable", when we are faced by our failures in the broader area of Information Management. The data element is the smallest piece of information in a total information system and yet we often seem to be placing our most intensive efforts into its "management"....ignoring its proper place in the scheme of providing intelligible, useful information to management or to other users.

With this in mind, we offer our comments in an attempt to relate this specialized field to its proper area and ask your consideration of the total aspect of Information Management.

2. CONCEPT

In this discussion of INFORMATION MANAGEMENT, we shall make several assumptions:

1 - Information is a RESOURCE, much as funds and manpower are, and should be managed as such;

2 - Proper balance should be maintained between the expenditures of funds and manpower used in the various phases of Information Management;

3 - Information that is NEEDED for proper management MUST be obtained regardless of cost;

4 - Information that is NOT needed for proper management, but is merely wanted, should NOT be obtained;

5 - Information that is NOT USED should be eliminated;

6 - The Information Management process consists of several phases resulting in an end product, a report, and involves a number of management disciplines including directives, forms, reports, records disposition, office equipment, as well as a number of others.

With these assumptions in mind, we shall briefly discuss the various phases in Information Management. These phases, in chronological order are:

Requirement
Acquisition
Manipulation
Dissemination
Storage and Retrieval

The Requirement Phase is that period where the need for the information is formulated and established. At this time, system design and development of the reporting system or requirement should be undertaken. All too often this activity is postponed until the Manipulation Phase and is treated as an afterthought--too late to serve its proper function. The requirement is assessed, need determined, cost-benefit analysis made, system developed, forms developed, certification and validation made, and Report Control Number assigned during this phase.

Two factors increase the difficulty of this phase. First, there is the normal tendency of ADP systems to "sell" their services to more and more customers in an effort to fully utilize their capacity. This ultimately leads to an overload of the computer necessitating additional capacity (usually a larger, later generation model) which leads to further need to sell...leading to another computer...leading...etc., etc. Second, and frequently a result of the first factor, is the conversion of programs when a new computer is installed. It is most important, at that time, that a thorough systems design be undertaken. Often, in the haste to become operational on a new computer, the old programs are used as is, or are converted on a one-for-one basis. This is the ideal time to thoroughly question and justify the needs and to ensure that only needed information is processed or retained and that all re-programming is made with this in mind. With a one-for-one conversion, it is possible that a third or fourth generation computer is making use of first generation "systems"....a truly inefficient situation.

During this Requirement Phase, the disciplines involved are: Directives, Forms, Reports.

We have attached a chart (Figure 1) which shows the various phases in the Information Management process and the different management disciplines which are associated with each phase. This should help in an understanding of the inter-relationships between the disciplines.

The Acquisition Phase is the period during which the information, or data, is being generated or collected to fulfill the requirement. It is also the phase which most strongly impacts the field installations and gives them the most concern.

The Manipulation Phase is the phase which involves the various operations such as collecting, assembling, displaying, formatting, re-formatting, ADP processing, printing and publishing, and other actions to prepare the information in its final format.

The Manipulation Phase is the single phase where the greatest amount of resources, both funds and manpower, is concentrated. This has been shown by the cost figures we have accumulated over the past several years in our own reports management program.

The Dissemination Phase is simply that phase during which distribution of the report is made.

The Storage and Retrieval Phase is that phase in which the report is filed, maintained, retrieved for use, and sometimes identifies other information needs.

All of the above-mentioned phases combine to form an information system. As an example of how this approach may help, let me cite from our experience.

Early in NASA's reports management program, we found one network of over 40 reports. These included ADP, manual, and mixed forms of reports which originated in the field installations and in four different levels in Headquarters. There was redundancy as well as the transmission of duplicate reports via alternate channels. We found that, initially, we were providing information to seven external agencies and, at times, rather low level internal organizations were providing the information directly to external requirers while similar--or identical--reports were being transmitted by a high level organization who had obtained the information from the lower level. The dangers inherent in such a situation are, we feel, readily apparent.

3. CONCLUSIONS

As will be noted (Figure 1), not all disciplines are applicable in every phase. Yet, each is important in its own right and serves a useful function contributing to the whole if careful attention is given to matters of organization and leadership which we shall discuss in a moment.

An audit report illuminated some problems within each discipline which needed attention. We believe that the Reports discipline has achieved significant results in its area and has reduced most of the problems discussed in the draft report. This achievement has been due, in large part, to the increased top management interest and support over the past twenty-six months. Other areas have undoubtedly achieved some results with more improvements to be realized from the support and interest of management. Grouping related information disciplines under a responsible official should lead to improved management of agency information resources. We believe that this would benefit by the institution of an INFORMATION MANAGEMENT activity as opposed to the more parochial concept of Paperwork Management--which term has fallen into some disfavor in recent years due to overzealous marketing of its magic.

Since reshuffling of billets does not necessarily solve problems, it is important that the individual selected to supervise a cohesive Information Management organization must be knowledgeable in the broad spectrum of management disciplines, courageous, and willing to work towards gaining and maintaining top management support for improved information systems throughout the agency. Above all, he must never allow an information management discipline to become a ritual--rather the information disciplines must always be carefully tuned and balanced to enhance the total information process.

We do not believe that any further internal study is necessary to determine the logical grouping of disciplines which are, for the most part, grouped in the same general fashion among the Federal Agencies and at our own installations. However, there are a few pertinent questions:

1 - At what level of management should the organizational responsibility be placed?

2 - Under which major office function should it be placed?

3 - Where should the Headquarters installation function be placed relative to the agency function?

In the next section, we offer some possible answers to these questions and suggestions for implementation.

4. SUGGESTED IMPLEMENTATION PLANS

With the foregoing points in mind, we offer the following proposed actions for your consideration:

1 - Accept and approve the Information Management concept;

2 - Identify specific elements involved and determine the division of effort for supporting Headquarters as an installation versus providing information systems leadership to the agency.

We will address ourselves here to providing some possible answers to the questions raised in the previous section.

In offering an answer to question one above, and considering the numbers of disciplines and people involved, we would recommend that a separate Office of Information Management be established. The function MUST receive attention from this level, as a minimum, if it is to gain proper top management support.

Consideration of question number two suggests that there are only a few organizations under which the function could be logically placed and the most logical of these, we believe, would be in the office of systems management or the office of administration, whichever held the responsibility for overall management information rather than that office which merely controlled the ADP portion of the information system.

The answer to the question of the placement of the Headquarters installation function (number 3 above) offers several possibilities.

We recommend that the functions be made a Division under the Director of Information Management. Some overhead savings should be realized by this placement. However, precautions should be taken so that (a) its priorities as an operating activity do not, in the event of personnel shortages, de-emphasize its Agency functional responsibilities, and (b) consideration is given to operating in a manner consistent with other Headquarters offices with dual responsibilities for the agency and for Headquarters as an installation.

3 - Select Information Management supervisor;

4 - Draft revised functional statements for each affected unit and submit for approval;

5 - Review and approve goals and objectives of Information Management unit.

Adoption of these suggestions with the support of top management, and the proper leadership, should lead to improved information systems.

5. SUMMARY

We believe: (1) that there is a need for concentration on INFORMATION MANAGEMENT as a whole, rather than emphasizing data element management; (2) that there are certain assumptions that must be made and that these assumptions are valid; (3) that there are several phases in Information Management and that each plays an inter-relating role with each of the others; moreover, certain of these phases make heavy demands upon resources; (4) that there is a need for organizing and implementing a total Information Management program in every organization; and (5) that the head of the Information Management function should be fearless, aggressive, and capable of obtaining and maintaining the support of top management.

Only through the application of such ideas can a total management information system survive and flourish--free of duplication, redundancies, and inefficiencies. We must step out from the relative restrictions of data element management into the wider world of Information Management.

DISTRIBUTION OF MANAGEMENT DISCIPLINES

PHASES DISCIPLINES	REQUIREMENT	ACQUISITION	MANIPULATION	DISSEMINATION	STORAGE AND RETRIEVAL
CORRESPONDENCE		☒	☒		
DIRECTIVES	☒				
FILES			☒		☒
FORMS	☒				
MAIL				☒	
MICRO-FILM					☒
OFFICE EQUIPMENT		☒	☒	☒	☒
OFFICE METHODS		☒	☒		☒
PRESENTATIONS		☒		☒	
PRINTING AND PUBLICATION				☒	
RECORDS DISPOSITION					☒
REPORTS	☒	☒	☒	☒	☒

NOTE: In each discipline above, it is assumed that the appropriate steps in management analysis are taken.

FIGURE 1.

ADDENDUM

The questions were most interesting and revealing. I have attempted to answer them fully. Should there be further questions or comments, I would welcome them in writing or via telephone. See Attendance section for address.

The first two questions below were answered at the Symposuim and the answers are taken verbatim from the tape of the meeting.

QUESTION: "'Why should the Office of Information Management be separate from the ADP Operations Office?'"

ANSWER: "The Office of Information Management should be separate from the ADP Operations Office simply because, in our experience, the ADP operations office tends to view the computer as a god and worships it...and everything is the computer...'the computer MUST have'...'this is the way it has to be'. Information Management cannot work on that premise. Information Management HAS to control the computer instead of having the computer control Information Management."

(NOTE: the applause following this statement was MOST gratifying. Thank you!)

QUESTION: "Two related questions: 'You say that information is a resource, often acquired at great expense, but that it should be discarded if not used. Why should it EVER be discarded? Don't you think the problems and costs of permanent retention reveal a technology lag? The second one is: 'Re: your principle of not obtaining data not directly needed...if pertinent peripheral data can be collected easily, doesn't this promote flexibility for users in defining requirements and producing impromptu reports?--i.e., helps build a powerful, flexible data base.'"

ANSWER: "Why collect it if you don't need it? This has too often been done in the past. Let me give you a couple of examples real quickly: we have just finished a World Series--also the Census Bureau. The Census was originally designed for allocating representatives...that was the reason for counting the people in the country. Now the Census form, which you are required by law to answer, asks how many bathrooms you have, what the construction (material) is of your house...What the hell does that have to do with how many make up a constituency? The second is, in the World Series we collect all these statistics...I call it statistics for the sake of the statisticians...I think if you had asked the question last night of the announcer, 'When was the last time a man with one green eye and one brown eye, batting left-handed, throwing right-handed, wearing white shoe laces, wearing a red cap and blue underwear came to bat?'...He could have answered! WHO CARES? If you don't need the information, don't collect it. If you do need the information, then get it and use it."

In reviewing the tape, I noted that I failed to completely answer these two related questions and should like to do so now.

You should discard unneeded information for several reasons: it costs money to store--by occupying computer, tape, card, or hard copy storage--and provides an opportunity to resurrect it easily, thus enhancing its possibility of survival...if it is dead--bury it!; retaining such information also tends to lead to the generation of additional reports--based on such data--which may be of little value. I feel that there is very little technology lag in the "problems and costs of permanent retention" as there are a variety of methods for storage especially in the areas of microfilm--what we seem to have instead is a technology lag in the proper design and utilization of information processes. Careful thought in information systems design can reduce the information required to the minimum and present the most economical method of obtaining and presenting it...this is not necessarily the computer. Systems development costs for computer-based information systems are high and we have already discussed the problems of computer replacement. This also applies to the matter of obtaining "pertinent peripheral data" where such activity is frequently undertaken for the benefit of job security or job enhancement. The matter of flexibility, impromptu reports, etc., should remind us that this activity provides an opportunity to provide a "shopping list" of data elements to a potential user...who promptly says, "I'll take six of those, a dozen of that, and...oh, yes...four each of those too." Ascertain needs...meet the needs...and keep costs to a minimum.

QUESTION: Please relate "Information Management" to or with the "Data Administration" function. Are they the same, in parallel, or what?

ANSWER: Information Management refers to the total function or process of providing needed information in a useful manner and on a timely basis. This information may be manual, computer derived, or a combination of manual and ADP. Key to Information Management are the words "needed, useful, and timely". Data Administration most frequently...in fact, nearly always...refers to that portion of Information Management which deals with automated data processing methods and often fails to examine manual systems or overlooks necessity or utility in its ADP operations due to the relative ease of manipulating data.

QUESTION: From your experience with "MIS" (Management Information Systems), what are the principal problems of the Data Manager, e. g., Management resistance, support, etc., when setting up a MIS?

ANSWER: Some of the important problems in establishing a MIS are the lack of management support, resistance to change, parochial interests, and lack of knowledge. Management often takes the process for granted and fails to make known its interest in a viable information system which provides for its needs in the most economical and timely manner...until the costs of systems become so great that management must intervene to reduce costs by eliminating unnecessary, untimely, and/or useless reports. Reports of marginal utility also receive attention at this time. Resistance to change occurs due to unfamiliarity with the new system, fear of job security, and failure to see the deficiencies (or failure to admit to these deficiencies) in the existing system. Parochial interests tend to require the continuance of a report or system due to "image puffery" (overstating the importance of one's own function), and job security. Lack of knowledge occurs at all levels...management simply may not know the most effective method of obtaining information and those involved in providing the information frequently are unwilling, or unable, to make alternate methods known to management because of some of the same factors mentioned in the first sentence: support, change, interests, etc.

QUESTION: What background do you suggest individuals should have in the Office of Information Management?

ANSWER: As to specific formal education, it matters very little. My own degree field was International Relations and I know of another Information Management persuasion individual whose degree field was mathematics. Primary requirements are a general knowledge, intelligence, and an inquiring mind. The individuals should be knowledgeable and skilled in management analysis...whether this is obtained through formal education or experience is relatively unimportant. Work experience in a variety of fields often provides a storehouse of information which can help in understanding problems and assists in analysis of them.

QUESTION: How closely should the/an information plan or data management be structured to the line organization of a company? What problems could there be if the information organization and line organization differ? Should a line organization change because of the logical information sources and flow?

ANSWER: The function of information management is to meet the information needs of an organization, whether line or staff. The structure that is necessary is that which meets the needs and, since needs do change from time to time, the structure must be capable of change also. The information management group serves in a staff function to management and is responsive to management's needs: this function may be by a separate group which serves all other groups in the total organization. It may also be a small group within a line organization in which case it must also be aware of the total process of information management throughout the organization and must avoid the parochial viewpoint frequently encountered in parts of the whole: in this type of organization it must make special effort to avoid duplication of information...an especially difficult matter in such a situation. I believe the information organization should differ from the line organization, i.e., the line organization should not be the information organization, but each does have its unique responsibilities within its own field...the line organization normally for a program or product and the information organization for an internal product--information--yet the information organization can only provide an effective and efficient product with the support and approval of top management. A line organization need not change its structure because of the logical information source or flow. It may, however, be made more efficient by a re-structuring and this must be determined by circumstances. Resistance will be met in this case for many of the reasons mentioned in the answer to the fourth question above. Information management function and organization should be flexible enough to meet information needs without organizational re-structuring. Again, circumstances in each individual organization will be the determining factor.

QUESTION: Are "information management systems" (in NASA), systems that produce reports that "advise" managers to confirm the "decisions" (made by computer) OR are reports (which are) produced used as supportive documentation for the "decision-making" process (by managers)?

ANSWER: They are used to provide information to assist management in its decision-making functions. We must face the fact that computers do not make "decisions...they may have been given certain parameters and told that if certain things occur the computer shall provide a certain response or course of action. In this case, the decision was made prior to the computation by the machine and the alternatives provided to it...the computers "decision" then is merely whether or not the information is within the parameters.

QUESTION: Is your goal Records Management? Forms Control? Library Centralization-Decentralization? Report Distribution Control? Paper and Microfilm Formats? Abstracts? Indexing? Term Definition?

ANSWER: YES! We mentioned the inter-relationships of a number of disciplines. The functions mentioned are disciplines or sub-elements of a discipline involved in the total information management process. Each has a role to play in the total process. In our own area, Abstracts, Indexing, and Library Centralization are normally associated with a Scientific and Technical Information Facility and are not used routinely in the management process. However, if Library here refers to ADP program libraries or a central location for the distribution of reports to management this function would be of concern to an information management group.

QUESTION: You implied in your talk that the information manager should be prepared to tell his boss (presumably a "user") whether or not he NEEDS certain information. Surely this is usurping the role of senior management to decide what he needs.

ANSWER: I think I stated it more strongly than that! It is not necessarily usurping senior management's role...it may be helping him to determine what he needs for I think we have all seen cases where management really doesn't know what it wants so asks for more than is really needed. The information management role is to help management to determine what is needed. Top management should have no, or little, need for detailed information as lower exhelons should provide their share of the management function and digest this information. Should top management insist on detailed information it may be due to ignorance, lack of proper management at lower levels, or mistrust of lower level managers. The cures here are not necessarily more information, but management action to overcome the deficiencies.

DATA STANDARDIZATION

Harry S. White, Jr.
Associate Director for ADP Standards
Institute for Computer Sciences and Technology
National Bureau of Standards
Washington, D.C. 20234

Data standardization is a subject that is often misunderstood. Perhaps this is because we have not yet begun to adequately recognize the significance of data as one of the major components of an information system. Up to now, data has been given a role of secondary importance in comparison to the hardware and software components of an information system.

Before proceeding further, we should first stop and examine the premise, "Is data a major component of an information system?" and "If it is, what should be our concerns about data?" Finally, we should address the question "What is the significance and impact of data standards?"

Is Data a Major Component of an Information System?

In order to address this question it is first necessary to assure that we understand what is being asked. There are two terms in the question "data" and "information system," that conceptually present difficulties due to the different perceptions each of us has about them. For purposes of this paper, definitions of these terms are provided in order to establish a framework for further discussion.

> Data - an ordered representation of letters, numbers, systems or other intelligible means used to convey meaning.
>
> Information System - The combined assembly of operations, processes and resources that are involved in the design, development, collection, operation, maintenance and use of an information service.

If we identify the resources that comprise an information system, we immediately are able to name: (1) the equipment, (2) the software, (3) the personnel, and (4) the facilities. Usually, we neglect to include the data in this listing of resources. Why do we overlook the most vital element of an information system? Are we unable to deal with data as we are other resources? Are we unable to measure data in terms of its value or quantity?

WHAT SHOULD WE BE CONCERNED ABOUT?

Until we are able to manage data like other resources, we will not be able to effectively control our information systems including their cost and efficiency of operation.

The computer community has produced new tools for data management, but for the most part these are directed toward data management at the micro level. Within the computer environment, data consists of raw materials to be processed, stored and manipulated. If we are to really manage data, we must first provide a much broader perspective than the computer view of data. There must be management controls for the collection of data, for the protection of data and for the processing of data. We need to establish procedures for getting rid of data that is no longer needed and presents both security and storage problems.

WHAT IS THE SIGNIFICANCE AND IMPACT OF DATA STANDARDS?

Data standards, as such, are not a new concept. We have been exposed to data standards in our everyday lives for many years. For example, the calendar is a form of data standard for identifying the days of a year. House numbers are a form of data standards for identifying different houses on the same street. Page numbers in a book are data standards used to identify different pages in the same book. With the increased use of computers, it has become essential to adopt data standards to provide for the effective processing and interchange of data. However, as might be expected, data standards have been developed for specific applications, usually at the installation or system level. As the exchange and collection of data has been extended to include machine readable data, we find that we cannot effectively communicate due to the differences in the definitions assigned to data or the differences in data coding. The primary purpose of data standards is to overcome these problems of incompatibilities, thus making the collection, exchange and processing of data more effective by eliminating unnecessary translations and improving communications.

In the Federal program for data standards our objective is to make maximum utilization of the data resources of the Federal Government and to avoid unnecessary duplications and incompatibilities in the authorized collection, processing and dissemination of data. Essentially this objective forms the basis for the voluntary standards activities of the American National Standards Institute and the International Organization for Standardization.

FACTORS THAT INHIBIT STANDARDIZATION

Based upon nearly ten years of experience in the development of data standards, I would like to reflect on some of the factors that inhibit or discourage the development of useful data standards. We must consider these factors, if we are to produce standards that are timely and responsive to user needs. I feel that the most significant factor that serves as an impediment to the development of data standards is the difficulty of being able to predict or demonstrate in concrete terms the savings, cost avoidance, or increased productivity resulting from data standards. We have not yet been able to come up with an acceptable method that can be used so that management can make the proper decisions in regard to supporting budget initiatives or placing priorities on continuing standards efforts. Contributing to this situation is the fact that in the past, data standardization has been undertaken in isolation and treated as a subject separate and apart from the data management function. Until we are able to provide adequate accounting mechanisms for our data resources, we will continue to have marginal support for data standards. In this regard, we must refocus our data standards efforts to be supportive of the data management functions. Otherwise, data standards in isolation will be searching for problems seeking a solution rather than providing solutions for identified urgent problems. Effective data management will enable us to pinpoint those subjects or areas where standards can be useful and also this will provide the basis for projecting the benefits and savings that will result from standards.

STANDARDIZATION MEANS COMPROMISE

Another major factor that has significantly reduced the effective development of standards is the recognition that the standards process by its very nature means that reasonable compromises must be made. It usually is not possible to adopt without change existing practices for use as general standards. The primary reason is that most existing practices have been developed as a solution for a specific application and need to be expanded or redefined before they are acceptable in other applications. Accordingly, compromises must be made before a standard surfaces.

It is difficult to contribute effectively to the development of a standard that is different from the method you have already implemented and will cost you to change. However, one must be able to make reasonable compromises if the long term need for the standard outweighs the cost of conversion. Reasonable compromise does not mean that the quality of the standard needs to be lowered, but does mean that each participant in the development of the standard needs to be able to be flexible and understanding of the needs of others. The standards participant must be able to consider alternative solutions and respond in a positive constructive way. Before appointing or selecting individuals to participate on standards groups and represent your organization, it is essential to discuss the matter of reasonable compromise and establish an understanding with the participant as to the manner in which he represents your interests and the acceptable results expected.

STANDARDS MUST BE TAILORED TO APPLICATION REQUIREMENTS

Another significant problem that has contributed to delay the development of standards is a misunderstanding on how the resulting standard is to be applied. Too often the standard is envisioned to be a universal solution for all applications. This approach has very practical limitations. For example, in the development of the general standard for coding organizations, it was recognized that the number of organizations to be coded numbered in the millions. This had to cover manufacturers, government activities, hospitals, retailers, libraries and all other entities that qualified as an organizational unit. Accordingly, the size of the code had to be large enough to cover the total number of organizations involved. This meant that the numeric code had to be at least eight digits in length in order to allow for additional code assignments.

The problem resulting from this solution is that most applications that need to code organizations are dealing with a subset or selected class of organizations. Usually the number of organizations in a particular application is less than the total number provided for in the general standard and can be coded using a lesser number of digits. In this situation, if the general standard were applied in each application, the overall operational costs would be increased. Alternatively, if the general standard were used in the interchanges among dissimilar applications, the overall operational and implementation costs would be reduced and interchange made possible.

This would then allow each application to develop and use standards fitted to their needs and to use the general standard as necessary when communicating with others outside the application area. We have learned that standards must be developed and applied in a realistic way appropriate to differing and changing environments. To standardize otherwise, is strictly arbitrary and unwarranted.

POSITIVE FACTORS

These essentially are the major factors that inhibit the production of effective data standards. In a more positive sense, we need to examine those factors that result in useful standards. The first and most important is the recognition of the need for a standard as a solution to a specified or particular problem. This need should be justified and expressed in terms of reduced costs, improved productivity, or cost avoidance. The time and cost for producing, implementing and maintaining the standard should be reliably predicted.

The second important factor is that those activities that will either use the standard or will be impacted by the standard are included in the standards making process. This does not mean that each activity must have representation on the standards development group, but does mean that each activity is extended the opportunity to participate either actively or be kept informed of the progress of the standards effort.

The third important factor is the selection of qualified competent individuals to develop the standard. In addition to being able to make reasonable compromises as mentioned earlier, the individual must be result oriented. The most effective standards have been produced by busy managers with adequate technical staff support. This approach results in standards that are timely and responsive to immediate operational requirements and does not become an academic exercise.

STANDARDIZATION STATUS

Within the Federal Information Standards Program, six data standards have been produced. These are standards for representing dates, states, counties, standard metropolitan statistical areas, congressional districts and countries. Policies and procedures for data standards have been issued in the code of Federal Regulations and are published as FIPS PUB 28. Arrangements have been made with the Civil Service Commission to undertake the development of program standards for data elements used in automated civilian personnel systems. Of particular significance to both data management and standardization is the recent establishment of a FIPS Task Group (Task Group 17) that is undertaking the development of standards and guidelines for data element directories. This effort is expected to result in a guide for initiating and maintaining a data element directory. It will be addressed to data management in general to include automated and manual systems as well as forms design and information services. Also the task group is developing the performance criteria for an automated data element directory system for use by data administrators.

Nationally, under the sponsorship of the American National Standards Institute, there are four approved voluntary industry standards for data elements. These include representations for dates, counties, states and time zones. Six others are in the final approval process and include standards for identification of organizations, named populated places, time, countries, points and metric and customary units. Also, ANSI has produced a technical guide for the development of data standards. This guide is due to be published in early 1976.

Internationally, standards have been produced for representing dates, time, metric units and time zones. Work is continuing on the development of standards for representing organizations, industries, commodities, occupations, sexes, blood types, points and mailing addresses. Also under study is the need for standards in the application of check characters to improve the reliability of data inputs.

Information concerning the availability of national and international standards may be obtained from the American National Standards Institute (ANSI), 1430 Broadway, New York, New York 10018. Information concerning FIPS Publications and their availability may be obtained by contacting the Office of ADP Standards Management, Institute for Computer Sciences and Technology, National Bureau of Standards, Washington, D.C. 20234.

Data Element Analysis and Use of a Relational Data Base Structure for Mapping Bibliographic and Numeric Data Bases

Martha E. Williams
Scott E. Preece
Sandra H. Rouse

Information Retrieval Research Laboratory
University of Illinois
Urbana, Illinois

The totality of available data base resources in the U.S. and other countries is extensive and provides a tremendous potential for individual bench scientists, information scientists, and researchers concerned with interdisciplinary problems. Unfortunately, there is currently no systematic method for interconnecting the data bases, conversion tables, and search systems, and there is no system that points out the links from one information source to another. The use of data base mapping would be of significant benefit for wide scale resource sharing, network design, and for efficient data base use by information centers. In addition, a data base map could be consulted by individual scientists to quickly determine the location of machine-readable sources of use to them.

The purpose of the NSF sponsored program discussed here is to develop a data base mapping model and search scheme (DBMMSS) to determine the feasibility of data base mapping. The data base map will make use of the relationships that exist between available: data bases; conversion tables and algorithms; information centers processing data bases; search software systems; and data base standards. The mapping files contain descriptive data about not only bibliographic data bases but also data-type data bases. The data base mapping system is designed to permit identification of optimal and potential routes from data base to data base, using intervening conversion tables or algorithms, if needed, in order to obtain desired data.

In order to demonstrate the feasibility of data base mapping, we are developing a mapping model and will test the model with a sample file. The sample file or test file will contain data about chemical data bases. The tasks required for establishing and testing the model and mapping concept include: acquisition of information about data bases; data analysis of data bases with emphasis on data element identification, description and tagging; design of model file structure; development of a test file; and testing of the map search and routing concept via rudimentary software.

The mapping system is implemented as a relational data base. Information describing various data bases and their contents is stored in the system as are descriptions of types of information, data base processing centers and data base producer information, and software systems. Data bases described include bibliographic and numeric data collections and conversion tables and algorithms. The various entities described in the system are then connected by relations. Thus, the connection(s) between any pair of entities can be explored by the basic relation operations, "join" and "project."

The relational data management system supporting the map was built to handle the specific needs of the project. A relatively complicated data structure has been implemented to provide flexibility. A relation is stored as a linked list whose members are sets of pointers to data items and are also part of other linked lists representing subrelations. The data items are stored as records containing a varying number of varying length fields.

At present the system is operational on a DEC-10. All programs in the system were written in SAIL, an ALGOL-like language developed for use in artificial intelligence research.

Key words: data base; data elements; mapping; microelements; relational data base structure; substructure searching.

1. Introduction

1.1. Problem Definition

Networking and resource sharing of data bases on a large scale is an eventuality that is not far off. But, there are several problems associated with data base networking. The problems are not those related to hardware interfaces or communications technology. The operational status of the hardware/communications field is evidenced by organizations such as TYMSHARE, ARPANET (Advanced Research Project Agency Network) and the newly announced commercial networking organizations, PCI (Packet Communications, Inc.) and Telenet Communications Corporation. The real problems or roadblocks standing in the way of resource sharing and use of data bases via networks are the data bases themselves. Almost every individual data base stands as a unique entity organized for unique purposes. There is very little commonality of organization, format, content, method of representation, etc. While individual data bases are organized within themselves and several data bases produced by one supplier may have common formats (this has become true only in the last few years), it is glaringly evident that there is no standardization that applies to the multiplicity of data bases in use today. Many may contain the same data element but the tagging, coding and nomenclature vary.

The totality of data base resources in this country is extensive and provides a tremendous potential for the individual bench scientist and for the information specialist. But, because of the lack of standardization, there is no predetermined method for interconnecting the data bases, search programs, and conversion programs. Theoretically, one should be able to identify the name of a desired chemical from a reference source and then obtain desired information about the compound, such as physical properties, toxic effects, handling requirements, etc., from a wide variety of machine-readable sources. This, however, cannot be easily accomplished because the name or code for the compound may be represented differently in different files, and because there is no system that points out the links from one information source to another.

1.2. Objective and Approach

We are developing a Data Base Mapping Model and Search Scheme (DBMMSS) in order to study the feasibility of data base mapping. The objective of data base mapping is to identify the linkages (a link is a common data element even though it may be tagged, coded or named differently in different data bases) between data bases, and to map routes for accessing required data bases and conversion systems, regardless of their location, in order to obtain desired data. The data base of data bases will contain information about: (1) bibliographic type data bases--particularily scientific and technical data bases; (2) data type data bases; (3) conversion table data bases; (4) information retrieval center data; (5) search software systems; (6) conversion algorithm software; and (7) standards for data bases.

In order to demonstrate the feasibility of data base mapping, we are developing a mapping model and testing the model with a sample file. The sample file or test file will contain data about numeric chemical data bases. Chemistry has been selected because the problems encountered in chemical data bases are so numerous and varied that it should provide a "worst case" situation, and thus is an excellent test vehicle for the model. The model will serve equally well in other disciplines.

Chemical data exists in many forms but there is no adequate system for tying the many files together so that output of one file is useable as input to another file. The situation is analogous to that of the 1920's prior to the establishment of standards for electrical equipment. At that time equipment used in one city could not be used in another. Now standards have been enforced throughout the world and even though both AC and DC power are used in different locations, converters are readily available.

In theory, one should be able to identify desired compounds, and obtain relevant information about them, for example, from Chemical Abstracts Services (CAS) data bases. Then, using the CAS Registry Number output, he should be able to access a National Cancer Institute (NCI) data base; and from there he should be able to access Food and Drug Administration (FDA) files or other files that contain linkages in common with the NCI file. By using common data element linkages one should be able to locate interfaces between data bases and where linkages between data bases do not exist, he should be able to identify the conversion tables or algorithms that may be used to obtain them.

A typical search in which a data base map would be very useful might be a question asking what chemicals have produced toxic effects on lampreys, and what effects this substance would have on fish and humans. A data base map search would indicate: (1) what data bases cover the initial question regarding which chemicals are lampreyicides; (2) where services to these data bases are available; and (3) what information is needed for the search. It also might indicate charges for search services and the availability of other services (such as photo-copying) offered by each of the information retrieval centers that process the data base(s) appropriate for answering the question.

The output of the first search might be chemical names, chemical structures, and CAS Registry Numbers. Armed with this information the file could tell the user what data bases are available and can be searched using output of the first search as input to the next search, etc.

If there were a specialized data base containing data on insecticides, pesticides, and herbicides, and if that data base also used CAS Registry Numbers associated with its compounds, the searcher could access the data base using the Registry Number output of the first search. This Registry Number data element is the link between the two files. If the user also wanted information on the effects of the compounds on humans, and if that data were contained in a Food and Drug Administration file which did not contain Registry Numbers, but did provide access to compounds via Wiswesser Line Notations (WLNs), the searcher would need to find a conversion table or algorithm for converting either the compound name data element or the Registry Number data element to WLNs.

It is this type of mapping scheme for identifying the appropriate series of data bases and conversion systems, and accessing them in the optimal sequence that we are modeling and testing. If the mapping model and search scheme is effective for chemistry, it should also be useful for other individual disciplines such as law where all cases are identified by a unique citation number (such as 35ILL ND358 for volume 35 of Illinois Reports, second series, page 358). This citation number is a data element that would be used not only in the Case Reports (whether they be official reporters or commercial reporters) but in textbooks, statute books, or journal articles. The citation number could also be referenced in the Shepherd Citations where other cases that cited the case in question can be located through the same data element--the citation number. A similar data element is used for statutes and regulations. Some of this material now exists in machine-readable form and other sources are being, or will be, put into machine-readable form.

1.3. Tasks and Status

In order to meet the objective of the study--to identify the linkages between data bases and map routes for accessing required data bases through conversion systems, regardless of their geographic location in order to obtain desired data--it has been necessary to carry out the three basic tasks of data acquisition, data analysis and software design and development. We have obtained data on approximately 220 bibliographic data bases, 60 numeric and conversion table data bases in chemistry, and 100 centers that process data bases. A detailed analysis of the microstructure of data elements has been carried out for approximately 60 data bases and software for data management and mapping has been designed and is being tested. Data for more than 100 bibliographic data bases have been entered into the test file. After testing the software the DBMMSS will be evaluated.

2. Acquisition and Analysis of Data Base and Data Elements Information

2.1. Acquisition

In order to obtain data for the data base mapping system--which actually constitutes a data base of data bases--we did the following: (1) obtained data, from data base producers and from the open literature, regarding bibliographic and numeric data bases and the software used for searching them; (2) obtained data, from the open literature and through telephone discussions with appropriate experts in the field of chemical data systems, regarding conversion programs, and conversion tables for converting various representations of chemical substances from one form to another (e.g., graphic structure to chemical name; chemical name to code, etc.); (3) obtained data, from centers that process data bases, regarding the data bases they process and the services they provide; and (4) analyzed documentation supplied by data base producers regarding the data element content and format and representation of those elements.

2.2. Analysis

Data obtained in (1), (2) and (3) above were entered on-line into the relational data base structure file described in section 3 of this paper. The data base mapping model requires the development and use of common data element descriptions and a standard tagging scheme so that the same type of element in each of the several data bases could be identified and interconnected for purposes of file manipulation. The data elements had to be directly addressable via their tags. Tags were applied to the data elements of the bibliographic, chemical-data, and conversion-table type databases. In recording these data we have assigned tag numbers to 353 separate types of information elements that relate to data bases, centers, programs, etc. Relations have been established between certain tagged items in order to display meaningful relationships in response to likely user questions. For example, a subject category will lead to a display of a list of data bases appropriate to the subject, or the selection of a data base will lead to a display of the names of centers that process that data base, etc. In the case of chemical data bases, where nomenclature is a real problem, one could enter a trade name of a compound plus an indication of the type of information desired, e.g., carcinogenic effects. The system would then indicate the map or route required to get to the data.

For example, when looking for carcinogenic effects of aspirin, where aspirin is a trade name, and if trade names are not used for the substances in the file that contains carcinogenic effects data, the DBMMSS will display the required map. It will indicate the name of the appropriate conversion tables and/or programs, or sequence of tables and programs, needed to obtain the name or code used for the substance (aspirin) as it is represented in the file that contains the carcinogenic data.

Mapping between and among data bases, data elements, centers, programs, conversion tables and algorithms, subject categories, etc. is done via the DBMMSS. The data element definitions and representations of each data base had to be analyzed and compared prior to determining the content and representations for the DBMMSS. Mapping at the microelement level has also been done. Mapping of data elements and microelements found in bibliographic data bases has been done manually, although if time and money permit they will also be entered into the DBMMSS for automatic mapping. For our purposes a microelement is considered to be the smallest meaningful unit of information in a data element. A macroelement is considered to be a generic category that groups together a number of related data elements.

Anyone who works with two or more externally generated data bases quickly learns that a major problem associated with comparing the content and treatment of data found in any two data bases is the lack of uniformity of content, representation, and format of data elements. This problem is compounded when one works with multiple data bases. In order to see the full range of variety it is necessary to analyze data bases in many different subject fields and to analyze different types of data bases--bibliographic, numeric and conversion table, etc. The number of variations found in bibliographic data bases is large. They vary with respect to: subject matter; data element content (author, title, journal names, subject code, date of publication, etc.); structure of the file (order of records in the file, blocking factor, etc.); format of the records (sequence of elements within the records); representation of elements (upper and lower case, use of abbreviations, use of codes, etc.); character representation (ASCII, EBCDIC, etc.); character set; and other physical characteristics of the tape.

We analyzed each of 60 data bases by breaking them down into their data elements and their component microelements (the smallest meaningful components or atomic parts). Each data element or sub-element was uniquely defined and associated with a tag or identifier. In addition to mapping the microelements into elements and macroelements, relevant standards of ISO, ANSI, NFAIS-FID, MARC, COSATI, and others were analyzed and mapped together with the data base information.

2.3. Mapping Microelements

Procedurally, we analyzed the 60 data bases recording the lowest level meaningful microelements, together with their associated format and representation information, for each particular data base. Using the full range of the elements we then, through a series of hierarchical steps mapped upwards until we developed a small number of macroelements that could be used for describing the multiple data bases.

The hierarchy consists of 10 levels. We refer to the highest level as the macro level. The 8 levels between micro and macro represent various elements and meaningful combinations of elements found on the different data bases. The reason for breaking all data bases into their microelements, before determining what the macroelements should be was the fact that different data base producers define elements in different ways and they include different sets of microelements in their elements. Some also group various sets of elements in fields and assign tags or specific positions to the fields rather than to the component elements. Unfortunately, the term "data element" is used very loosely and has different meanings for different data bases.

At the microelement level we defined 85 microelements and at the macroelement level we defined six macroelements. The number of intervening steps between a microelement and its appearance in a macroelement varies considerably depending on: the type of macro; the data base concerned; and the complexity of the information represented.

The six macroelements for bibliographic data bases are: (1) secondary source information; (2) identification information; (3) bibliographic description; (4) intellectual content information; (5) availability information, and (6) processing information. The microelements are too numerous to list but a few examples are shown in figures 1-3.

Figures 1-3 show the variety of representations and formats found in three data elements in eight data bases. Figure 1 shows the variety found in "Personal Names;" figure 2 shows the variety found in "Corporate Authors or Patent Assignees;" and figure 3 shows the variety found in "Codes for Classification."

Figure 1 shows the variability found in the "Personal Name" data element. That element is comprised of four microelements and they are designated: (A) surname; (B) first name or initial: (C) second name or initial; and (D) post particle. In figure 1 the first column indicates the name of the data base, the second column indicates the application of any one of several national, international, or widely used internal standards. Where an internal standard has been used by a single data base producer for one or two of his own data bases, that "standard" is not indicated. Column 3 indicates the tag data and/or position information employed by the data base in question to designate the "Personal Name" element. Where a **>** symbol is used, this means that the data base producer tag specified includes in it more data than that which is associated with the data element "Personal Name." Column 4 indicates the name that the data base producer associates with the tag for the element. Column 5 indicates storage mode. Column 6 indicates which of the microelements we define as being a part of the "Personal Name" element are found in the element in the examples given by the data base producer. Where microelement designators are not shown one cannot infer that the microelement does not exist in the data base. We simply did not have sufficient information to make the positive determination. Column 7 provides information regarding multiple entries for the element. Examples showing the way personal name data are presented on the data bases are given in italics. Examples are given only if we were able to find them in the data base producer documentation.

Figure 2 indicates the variety of formats found in the "Corporate Author/Patent Assignee" data element. That element is comprised of three microelements designated as: (D) Parent Organization; (E) Subdivision or Part Identification, and (F) Code Identification. The explanations of the seven columns and examples are the same as for figure 1.

Figure 3 indicates the variety of formats found in the "Codes for Classification" element. In this case each specific code is a microelement as subject classification codes are specific to individual data bases. Explanations of the columns and examples are the same as for figures 1 and 2.

2.4. Implications

This microelement analysis for building generic macroelements does indicate the ability to map all important elements found on the commercially available machine-readable data bases to a standard structure regardless of the many different ways elements are named, tagged, or grouped on different data bases. The analysis was done on the bibliographic data bases. We are now in the process of doing the same thing for chemical data type data bases.

Once completed, appropriate data element data will be entered into the data base mapping system together with the relational information so that ultimately we will be able to display relational information; associated with data elements across and within data bases; associated with data bases and data base processing centers; and associated with data elements and conversion algorithms and conversion tables. Linkages between data bases, or between data bases through conversion algorithms and tables, are indicated by virtue of commonality of data elements. The data element analysis described in this paper is one of several tasks required to develop a DBMMSS and hopefully will contribute toward the goal of networking of multiple data bases in many locations throughout the country.

3. Relational Data Base

3.1. Design Requirements

The basic needs for the system were the ability to store and retrieve information about data bases and to build "maps" to guide the user from one source of information to another.

Experimental use of the pilot system convinced us of the need to allow the desired exploration of the user: data connection problem required a vastly more flexible system, much richer in interconnections. For this ability we turned to Codd's relational data base structure. The resultant information system provides the ability to store and retrieve complex information records containing a variable number of variable-length records of one or more types. With this degree of flexibility we have been able to meet the needs of the project.

3.2. Hardware Environment

The DBMMSS support software is written in the SAIL language on the Digital Equipment Corporation (DEC) System-10 time sharing system at the Coordinated Science Laboratory at the University of Illinois. SAIL is a superset of ALGOL-60 built at the Stanford Artificial Intelligence Lab and contains useful constructs for dealing with strings and special data structures intended for use in artificial intelligence research. The I/O structure of the DEC-10 is available essentially as it would be to an assembly language program. The DEC-10 in use at the Coordinated Science Laboratory is a test site for DEC operating systems, and currently runs normally under the 602 level, virtual memory monitor. The system has 270K of core memory and uses both RP03 and RP04-type disk units.

3.3. Surface Structure

Data stored in the system is divided into classes. This first, broad level of division is provided to allow access to all the entities of a specific classification. In practice these consist of such classes as "data bases," data elements," and "information processing centers." Currently about 10 classes of data are in use. Data classes are referred to by unique names.

The members of a data class are called records. A record is, in general, the collection of data describing a single instance of the data class. A record is referred to by a unique name, assigned by the user when entering it. More than one name may be assigned to a record by the use of an alias (synonym). References to aliases by a user are transparently directed to the aliased record through a synonym dictionary. Names need only be unique within a single data class.

A record is a set of fields. Each field is a text string. Associated with each field of a record is a six character code, generally used to distinguish the context of the data. The code "DATE," for instance, might be used to indicate that a field contains the date of the information in the record. A mechanism is provided for associating codes with longer strings, so that more descriptive names of field types can be used in communicating with the user.

The mechanism for specifying connections among records is the relation. A relation consists of a set of ordered n-tuples, each of which specifies that the described relation is valid for those records. A typical order two relation is "is-processed-by," which consists of ordered pairs, the first member of which is a data base and the second a processing center. Relations may be of any order up to 62, and are defined in terms of the data class associated with each position of the n-tuples. The example just given, for instance, is defined as being an order two relation with a data base item in the first position of each ordered pair and a processing center in the second.

For convenience of access each relation is accessible also through its subrelations, which are defined as consisting of all members of the relation containing a given value in a given location. This permits finding quickly, for instance, all data bases processed by a particular processing center, by requesting that subrelation of "is-processed-by" defined by a position in the members and a value for that position. This causes substantial extra processing during file setup and maintenance, but was justified because it is the most common access pattern. In fact it makes the relation operations "join" and "project" easier to compute.

3.4. Implementation

The basic concept underlying the file structure is that of linked lists of linked lists. The information content is divided into two files, one of data--one or more classes of data records--and one of relations. This division is made for the reduction of disk contention and to permit better localization of data access. Two additional files, in the same format as the data file, are used for support functions: one contains the text of system messages, the other is used for translating between names and codes, used for data class names, data field names, and relation names. The help message file is for system support, but could be switched, for instance, to permit the use of different messages for experienced and novice users. The name file would ordinarily correspond to all the sets of data and relation files in a given area, but multiple, switched name files might be provided to allow, for instance, printing different field names during data entry and data display. Because pointers are to absolute addresses, only one relation file would normally correspond to a given data file, and vice versa.

At the head of the data file is a list of hash tables, each representing a data class. Each entry in a hash table points to a list of records whose names have the appropriate hash value. Each hash table is identified by a six character code representing the data class.

Each record is headed by a block containing the record's name and pointers to the next record in the same hash class, the first data field of the record, the list of subrelations containing the record, and the list of aliases for the record.

A data field consists of a text string, a six character code identifying the field type, and a pointer to the next field of the record.

An alias is the same as the header block of a regular record, except it contains a code identifying it as an alias, a pointer to the record for which it is an alias, and a pointer to the next alias for the same record. The aliases are chained together so that they can be changed if file maintenance moves the aliased record.

The relation file is headed by a list of relation definitions. Each definition is identified by a six character code, and contains pointers to the hash tables (in the data file) for the data classes corresponding to each position of the members of the relation as well as a pointer to the first member of the relation.

Each member of a relation consists of a pointer to the next member of the relation, a pointer to each of the records making up the member, and, for each of those, a pointer to the next member of the relation in the specific subrelation corresponding to that record appearing in that position.

Each record in the data file points to a list of the subrelations defined by that record's occurrence in a particular position in a particular relation. Each subrelation header points at the first member of that subrelation and at the definition block for the relation it is part of.

3.5. Implementation Notes

Clearly the wide use of pointers to connect fields, relations, relation members, and so forth means that the program will spend a lot of time chasing down chains of items. This could slow response time significantly. To avoid this the program attempts to put adjacent items of lists in the same block of file storage. Thus wherever possible the chain scanning will not require more disk accesses than necessary. To implement this, each request for space in the file specifies a preferred location in the file. As a result items are usually placed at the head of chains, so that the value of the list head pointer specifies the block in which the next item should be placed.

A uniform procedure is used for allocating and freeing space in the disk files of all types. If storage is not available in the requested block, or no block is specified, a new block is allocated and the storage allocated in it. Within blocks a first-fit allocation method is used. Continuous recombination of freed storage is used. Since speed of chain scanning was judged more important than storage utilization efficiency for our needs, no data-moving garbage collection is used.

Accesses to records in data files, both reading and writing, are performed by common subroutines. These provide for reading or writing a single word, a block of contiguous words, or a whole field. Buffers for the files are maintained in a list in order of last access. When an access request is received for a block not currently in core, the least recently used buffer is found and the block is read into it. Provision of a reasonable number of buffers also aids in cutting list accesses during chain scanning, thanks to localization provided by the allocation schemes. At present all of our programs allow user access to the current statistics on duration and frequency of use of the buffers.

The access subroutines form a hierarchy, the lowest level routines being used by those "above" them. The lowest level routines are used to access a single word of the data file or a contiguous block of words, or the test part of a field, or all the stored parts of a field. Next above those are routines that deal with the logical subparts of fields or of members of relations. On the next level it is possible to retrieve the members of a relation, or the records of a data class, one at a time via the "get-next" function. Finally there are routines that manipulate all the members of a relation, or all records of a data class, together.

Because the routines that function on the highest level use the "get-next" function to retrieve items one at a time, we can have relations that are defined in terms of other relations and not stored at all. These "pipeline" relations conserve space and are much more efficient than stored relations when only the result of a series of operations is of interest. Of course a SAVE command is available to store defined relations when desired (if a relation is to be traversed several times it is more efficient to store it).

An index is kept, for each data class, of the fields appearing in the class's records. This permits a specific field of a specific record to be found without scanning the list of that record's fields. Also, all instances of a specific field type may conveniently be found.

Pointers consist of a single 36-bit word. The upper half specifies a block number, the lower specifies a word offset in that block. The DEC-10 operating system restricts us to blocks of 128 words. Characters are stored in ASCII seven bit representation except for codes (field, relation, and data class) which are stored in SIXBIT to allow for an additional character. Because of the 128 word per block limit we have accepted a limit of 600 characters for field values rather than add a mechanism for connecting sections of a string.

3.6. Capabilities

As an experimental tool, the system exists in pieces. Currently separate programs are used for file creation, file maintenance, relation operations, and the user interface. These operate much like typical time sharing text editors, using single-character commands and prompting for necessary parameters. Basic help information is available in all modules, but in as much as this is basically a feasibility determination project, no comprehensive tutorials have been created.

For file creation a program is used that builds one record at a time. The data entry specialist enters a data field code and the field's value repetitively until the record is complete, then indicates record completion and proceeds to name a new record. Existing fields may be replaced by new values, but no editing capability is supported. This program is intended for use by an inexperienced data entry specialist, rather than a programmer.

For the correction of existing data and for the creation of data classes, a powerful editor program is available. This program, intended for the experienced user, is modeled on the SOS text editor in use on the DEC-10. Both record level (create, rename, etc.) and intra-field string editor (insert text, move pointer in string, etc.) commands are available in the editor.

The relations management program is used to define and delete relations, to add or delete members of relations, and to perform the basic relation operations join, project, and combine.

The user interface program provides access to the data base and its relational presentation for the end user. Rather than implementing an access language, such as SEQUEL for user access we have built a system that supports a range of access patterns. The user selects an access pattern and is then guided by the system in supplying the necesary parameters for his particular request. Since the specific thrust of the project involves only a few capabilities of the relational model, used in particular ways, it was more efficient to avoid the additional complexity of parsing a command language. The user interface includes drivers for displaying the contents of relations and records, for selecting members or records via Boolean combinations of condition tests, for analyzing the contents of data fields, and for producing the connection maps that are the prime subject of the research program. The interface program makes use of the same lower-level field, record, class, and relation operators that also underly the basic creation and maintenance programs, but a single user command or prompt sequence may involve many processing steps.

Additional "application" programs are used for generating special listings, analysing the storage efficiency of files, producing user tools based on the files, and for other purposes. These use the subroutines provided by the system for data access and buffer management.

4. Summary

The Data Base Mapping Model and Search Scheme project was carried out at the University of Illinois to determine the feasibility of a data base map for showing existing and potential relationships between bibliographic and chemical numeric data bases, and between data bases and conversion tables and conversion algorithms. Relationships exist by virtue of commonality of data elements in various sources. In order to define elements and establish a set of tags for denoting the elements a large number of data bases were analyzed and broken down into their component elements and microelements. The resulting set of microelements was used to build upwards to a common set of macroelements. The data element analysis was necessary in order to define the element tags for establishing relationships between identical and similar element types that may be represented differently in different data bases.

The DBMMSS has been implemented as a relational data base. The DBMMSS support software provides for flexible organization of and access to files of data and of interconnections among data. It provides a support tool for a program of research into the provision of guidance to the data base user in making maximum use of multiple data bases of various types.

As a research tool, secondary priority has been given to high efficiency, highest priority has been given to providing maximum flexibility. The result, rather than a polished, user-oriented information, is a non-centralized, flexible, powerful tool for exploring the needs of the data base user in identifying data resources--data bases, data base processing centers, conversion algorithms and tables--and relating one to another according to his own definition of a problem.

The goal of resource sharing is recognized by most scientists, information specialists, librarians, and other cost conscious persons as a worthy and necessary objective. However, achieving the goal is not an easy matter. We firmly believe that the outcome of this study will assist in paving the way towards network sharing of data bases.

5. References

American National Standard for Bibliographic Information Interchange on Magnetic Tape. American National Standards Institute, Washington, D.C., 1971. (ANSI Z39.2)

Astrahan, M.M. and D.D. Chamberlin. Implementation of a Structured English Query Language. Communications of the Association for Computing Machinery, 18:10, 580.

Codd, E.F. A Relational Model of Data in Large Shared Data Banks. Communications of the Association for Computing Machinery, 13:6, 377.

Data Element Definitions for Secondary Services. National Federation of Science Abstracting and Indexing Services, Philadelphia, Pennsylvania, June 1971.

Elcheson, Dennis R. A Correlation of Bibliographic Data Elements for Use in a Generalized File Management System. Journal of the American Society for Information Science, 24:1, 45-53.

McEwen, Hazel E., editor. Management of Data Elements in Information Processing. Proceedings of the First National Symposium, 1974, January 24 and 25. National Bureau of Standards, Washington, D.C., April 1974.

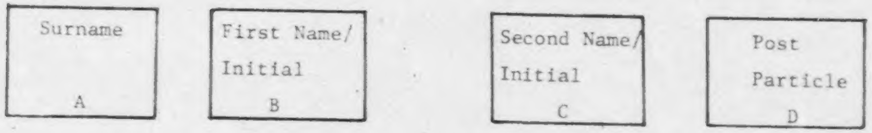
Schipma, P.B., M.E. Williams, A.L. Shafton. Comparison of Document Data Bases. Journal of the American Society for Information Science, 22:5, 1971, 326-332.

Schneider, John H., Marvin Gechman, Stephen E. Furth, editors. Survey of Commercially Available Computer-Readable Bibliographic Data Bases. American Society for Information Science, Washington, D.C., January 1973.

Williams, M.E. and A.K. Stewart. ASIDIC Survey of Information Center Services. Association of Scientific Information Dissemination Centers, Chicago, Illinois, 1972.

Personal Name

Author, Editor etc.



Data Base	Standard	Tag/Position	Producer Tag Name	Storage Mode	Micro-Elements	Multiple Entries
INSPEC	ANSI-MARC	200	Personal Author, Inv.	BCD EBCDIC	A-B-C-D	
		EXAMPLE: 1ƀDORNBERGER,J.F.,ƀJR.ǂ				
		EXAMPLE: 1ƀTEHƀFUƀYENƀDORNBERGER,ƀJ.F.,ƀJR.ǂ				
		210	Editor			Format for Multiple Entries is same as for Personal Author
		EXAMPLE: (Same Format)				
		220	Translator of Book			Format for Multiple Entries is same as for Personal Author
		EXAMPLE: (Same Format)				
MARC	ANSI-MARC	100 (with subf.a,c,d)	Main Entry-Personal Name	ASCII-8	A-B-C-D	
		EXAMPLE: 10ǂaKames,ƀHenryƀHome,ǂcLord,ǂd1696-1782				
		700 (with subf.)	Added Entry-Personal Name			Format for Multiple Entries is same as for Main Entry-Personal Name
PATELI		Directory Fields 1-4, Pos.32-47	Authors	EBCDIC		
		EXAMPLE: JamesƀA.Bonbright				
SPIN		*AUT (with subf.)	Author(s)	EBCDIC	A-B-C-D	Unlimited number of Authors
		*EXAMPLE: *AUTƀ¢AUIƀ%AUFƀJohnƀQ.ƀ%AUSƀSmithƀ%AUPƀJr.*				

Figure 1. Variety of Formats for the Personal Name Data Element in Eight Data Bases

Personal Name

Author, Editor etc.

Surname	First Name/ Initial	Second Name/ Initial	Post Particle
A	B	C	D

Data Base	Standard	Tag/Position	Producer Tag Name	Storage Mode	Micro-Elements	Multiple Entries
CACon	SDF	005901-0A	Personal Author Name	ASCII-8	A-B-C-D	Maximum 10 Authors. If more than 10, 9th is tagged, 10th tag is entered with last name "Other".
		EXAMPLE: Doe,ƀF.ƀD.,ƀJr.				
		007101-0A >	Patent Assignee	ASCII-8	A-B-C-D	
		EXAMPLE: (Same as 005901-0A in cases where assigned to person, company name also included)				Format for multiple entries is same as for Personal Author.
COMPENDEX	ANSI-COSATI	20	Author	BCD	A-B-C-D	Maximum 16 Authors
		EXAMPLE: SmithƀJr.,ƀJ.Q.|				
		EXAMPLE: Rassam,ƀPaulƀR.ƀEllis,ƀRaymondƀH.|				
GRA	ANSI-COSATI	280ƀ10	Personal Authors	ASCII-8	A-B-C-D	Maximum 5 Authors
		EXAMPLE: [J]ones,ƀ[J]ohnƀ[J]ƀR.				
		EXAMPLE: [J]ones,ƀ[J]ohnƀ[J]ƀR./[S]mith,ƀ[S]am.				
INFORM		201 (80 char. card images)	Author Data	EBCDIC	A-B-C	
		EXAMPLE: NEKVASIL,ƀCHARLESƀA.				

Figure 1. Variety of Formats for the Personal Name Data Element in Eight Data Bases--cont'd.

Corporate Author/
Patent Assignee

Parent Organization
D

Subdivision or Part Identif.
E

Code Identification
F

Data Base	Standard	Tag/Position	Producer Tag Name	Storage Mode	Micro-Elements	Multiple Entries
CACon	SDF	0059ƀ0B-14	Corp. Author Name	ASCII-8	D-E	Maximum 10 Corporate Authors
		EXAMPLE: Grace,ƀW.ƀR.,ƀandƀCo.				
		0071ƀ01-0A	Patent Assignee	ASCII-8	D-E	Format for multiple entries is same as for Corporate Authors.
		EXAMPLE: IUPAC-IUBƀComm.ƀonƀBiochem.ƀNomenclature				
COMPENDEX	ANSI-COSATI	20	Author	BCD		Maximum 16 Authors
GRA	ANSI-COSATI	300ƀ05 >	Corporate Source	ASCII-8	D-E	Maximum 400 Char.
		EXAMPLE: ↑B↓ureauƀofƀ↑M↓ines,ƀ↑W↓ashington,ƀ↑D.ƀC.↓				
INFORM		201	Author Data	EBCDIC		

Figure 2. Variety of Formats for the Corporate Author/Patent Assignee Data Element in Eight Data Bases

Corporate Author/
Patent Assignee

Parent Organization
D

Subdivision or Part Identif.
E

Code Identification
F

Data Base	Standard	Tag/Position	Producer Tag Name	Storage Mode	Micro-Elements	Multiple Entries
INSPEC	ANSI-MARC	730	Patent Assignee	BCD EBCDIC	D	Variable Length, Free Format, Subfields
		EXAMPLE: 1ƀGECƀLTD‡				
		EXAMPLE: 1$ICLƀLTD.$IBMƀLTD.‡				
MARC	ANSI-MARC	110 (with subf.a,b,t.)	Main Entry Corporate Name	ASCII-8	D-E	
		EXAMPLE: 10$aU.S.$ƀDept.ƀofƀState.$tTheƀDepartmentƀofƀStateƀBulletin.				
		710	Added Entry Corporate Name			Format for multiple entries is same as for Main Entry.
PATELL		Directory Fields 1-4, Pos.32-47	Authors	EBCDIC		
		EXAMPLE: Dinsmoor,ƀJamesƀA.				
		EXAMPLE: Dinsmoor,ƀJamesƀA.Bonbright,ƀJamesƀC.Lilie,ƀDanielƀR.				
SPIN		*AUT (with subf.)	Author(s)-Corp. Author	EBCDIC	D-E	
		*EXAMPLE: *AUTƀ*AUCƀBrookhavenƀBubbleƀChamberƀGroup*				

Figure 2. Variety of Formats for the Corporate Author/Patent Assignee Data Element in Eight Data Bases--cont'd

CODES FOR CLASSIFICATION

Data Base	Standard	Tag/Position	Producer Tag Name	*Examples*
CACon	SDF	031Bƀ00	CA Section Cross-Reference(s)	
		0067ƀ01	CA Publication Section/Subsection	*CA004003*
COMPENDEX	ANSI-COSATI	60	CARD-A-LERT Codes	*00-A406ƀ00-A411ƀ00-A412*
GRA	ANSI-COSATI	350ƀ02	COSATI/WGA Subject Code	[F]*IELDƀƀ9*[*A*],*ƀ1ƀ*[*B*].
INFORM	No Subject/Indexing Data			
INSPEC	ANSI-MARC	120	Sectional Classification Codes	*2$A1840ǂ;2$B3170$B1270ǂ*
		121	Unified Classification Codes	*1$NEGAABǂ;1$FECCAH$FCRAAX$WMEEAQǂ*
MARC	ANSI-MARC	050 (with subf. a,b)	Library of Congress Call Number	*0ƀ$aJK609$b.M2*
		082 (with subf. a)	Dewey Decimal Classification No.	*ƀƀ$a345.5/57$aB$a920*
PATELL		Directory Pos. 77-84	Classification Code	*06* *2 Digit Codes* *Maximum 4 Codes* *Prime Subject Content, Headings in PA Index*
		Directory Pos. 87-91	Subject Code #1	*15300* *5 Digit Codes* *Subject Index Headings for PA Index*
SPIN		*DAN (with subf.)	Document Analysis Information	

Figure 3. Variety of Formats for the Codes for Classification Data Element in Eight Data Bases

Status of the Army Materiel Command's Progression from Reports Control to Data Element Management

Edith F. Young

Reports Management Branch
US Army Materiel Command
Alexandria, Virginia

At last year's synposium, the Army Materiel Command's (AMC) plan to progress from a Reports Control to a Data Element Management Program was presented. In brief, the program first consisted of requiring a request form for a new manual report, assignment of a control number, and a yearly review for need of the report. Our inventory of controlled reports, mainly manual reports, has been completed with correlated preparation costs. It contains 935 reports costing over $10 million. Using this inventory as a base, controlled reports costing $1 million to prepare were cancelled during the preceding year. The second phase was the addition of all automated products to the program. Our inventory of automated products has also been completed. It contains 17,419 products costing almost $46 million at the beginning of FY 75. During the preceding year, products costing $8 million were cancelled as a result of command emphasis on the program. The third phase was to actually start managing, in lieu of merely controlling by the establishment of an AMC Data Element Standardization Program. We are approximately one to two years ahead of schedule, as we will have our first data element matrix analysis this fiscal year.

Key words: Reports control; data element management; Data Element Dictionary (DED); data element standardization; data element matrix analysis; data element management base files; data element characteristics; Army Materiel Command (AMC).

1. Introduction

At last year's symposium, the Army Materiel Command's (AMC) plan to progress from a Reports Control to a Data Element Management Program was presented. In brief, the program first consisted of requiring a request form for a new manual report, assignment of a control number, and a yearly review for need of the report. The second phase was the addition of all automated products to the program. The third phase was to actually start managing, in lieu of merely controlling. The term "managing" to us meant the ability to know the quantity of our data, to recognize unnecessary redundancy of data and to concentrate our resources to manage data in the high payoff areas. This required the identification and standardization of AMC information (data elements) and was to be accomplished by the establishment of a Data Element Management Program. A target of five years was set for the program to be initiated.

2. Data Element Management Philosophy

The AMC philosophy behind of this program was that all resources should be managed to insure our capability of "doing more with less" in these austere times. We felt that resources primarily consist of three facets; manpower, equipment and data. As a general rule, government and "big business" concern themselves with manpower and equipment expenditures, but seldom with data. One of the principle reasons for this is that an inventory with associated costs, has been difficult to obtain--there is no stereotyped accepted formula for identification and costing of this resource. Manpower and machine inventories/costs are readily available. However, when a comparison analysis was completed in AMC of the costs of these three resources, it was found, in many cases, the data resource consumes more of our budget to obtain, maintain and generate than either manpower or equipment.

3. Data Inventory

As stated previously, in order to manage our data, the quantity of our current inventory of data had to be identified. A follow-on action was also required to insure that only necessary, not nice to have, information was added to that inventory in the future. Costs, where possible to obtain, were also felt to be a vital adjunct to this inventory if a meaningful economic analysis of worth versus need were to be made. The inventory was stratified into two major components, location of data and identification/standardization of data.

3.1 Location of Data

Data is primarily located in three areas: forms, files, reports or products. AMC, based on our Reports Management Program, already had an inventory of controlled reports. It was therefore decided to complete our inventory of reports and ADP products first, then add forms and files (to include systems and installation) location of data.

3.1.1 Controlled Reports

Our inventory of controlled reports, mainly manual reports, has been completed with correlated preparation costs. It contains 935 reports costing over $10 million. Using this inventory as a base, controlled reports costing $1 million to prepare were cancelled during the preceding year.

3.1.2 ADP Output Products

Our inventory of automated products has also been completed. It contained 17,419 products costing almost $46 million at the beginning of FY 75. During the preceding year, products costing $8 million were cancelled as a result of command emphasis on the program.

3.1.3 Automation of Location

Systems analysis has been completed to automate the reports data location inventory, and programming is almost complete. We are currently loading our data base, with production targeted for January 1976. The master file in the system is called the Reports Attribute File (RAF) and will contain the AMC inventory of reports and ADP products. The RAF will contain one record for each report and ADP product; the record will contain attributes of the report/product. These attributes will include report/product title, costs, frequency, media, proponent, producer, volume (page count, frame count, etc.), and feeder information. The record will also contain the data element mnemonic (abbreviation) for each field of data in the report/product. The file will furnished management information on our reports and products requirements, such as the inventory and costs by report, ADP product, system, functional areas, producer, requirer, trend analysis. The system has the capability of furnishing both triggered and inquiry information. It will also enable us to identify the extent and specific data each report or product contains.

4. Identification of Data

One of the principle hinderances in installing the data element program has been the finite identification of data. Too often reports are approved which ask for data in generalities, such as "manpower data," not specifics such as "number of developmental systems analysts manhours." We are now enforcing the requirement that no new ADP systems/products or reports will be approved without a list of defined data elements. This still leaves us in the position of identifying the data elements in our current inventory of reports/products, which we have found to be no easy task.

5. Standardization of Data

Duplication and unnecessary redundancy of data have always plagued information managers. This has been like the weather; everyone complains about it but no one does anything about it. We in AMC are attempting to change this through our Data Element Management Program. This program was planned to accomplish two objectives, locate information and eliminate unnecessary redundancy of information. We found, however, before we could establish a program to accomplish our objectives there had to be proper communication and understanding of our information in reports and files. This required standardization of the data elements attributes, such as title, mnemonic, field length, and most important for all, definition.

5.1 AMC Data Element Dictionary

We have an active data element standardization program, with our final efforts being incorporated in our AMC Data Element Dictionary (DED). We feel we have an excellent DED software package system which is somewhat borne out by Department of Army (DA) having fielded our system as the DA standard; it is also one of the DED systems being looked at by the Bureau of Standards in its DED efforts. The dictionary has approximately eight thousand data elements. It contains the attributes of the data elements (name, definition, abbreviation, etc.) as well as the location of the data (files, forms, reports, products, systems, etc.).

6. AMC Data Element Management Program

Out AMC Data Element Management Program will use these two files and systems, RAF and DEDS, as its basic tools to manage data elements. The RAF will furnish the information to the DEDS as to the initiation, changed or deleted report/product requirements and location of data elements. The DEDS will furnish the range of locations for data elements or the lack of available information.

6.1 Location of Information

If a customer knows the name of the data element for which he wants information, the DEDS will now furnish him this information. The system will furnish him the attributes and various locations of the data element. We are currently attempting to develop an automated system for identifying the data element for the customer when he does not know the specifc name of the data element. This system is also planned to prevent duplicate standardization action on the same data element and to eliminate synonyms.

6.2 Redundancy of Information

It is anticipated the major benefits from the Data Element Management Program will be in the prevention and elimination of unnecessary redundancy of information within AMC. All redundancy is not bad; there are many instances of where it is required. We want the capability, however, of being aware where redundance is resulting in unnecessary expenditures of our resources. This system will enable us to know when data is duplicated in our computer files, systems, input and output products, as well as our manual reports and forms. It will also tell us when data is no longer needed so the data can be eliminated. The expense of generating data from a computer system into products is minor compared to the costs of gathering and maintaining that data. It has been evidenced, however, in most large integrated computer systems that there is no knowledge of when a field of data is no longer required. An output product can be cancelled but the source fields of data are not eliminated from the files based on the lack of awareness of all usages of the data, which results in not knowing when there is no usage.

7. Current Status of Program

It was originally estimated, in 1973, the AMC Data Element Management Program would require five years. We are now two years into the program, with three years to go. We are approximately one to two years ahead of schedule, as we will have our first data element matrix analysis output this fiscal year. The programs have been designed, programmed and tested. The data base to produce data element identification and location for the RAF is now being loaded, which will feed the location data to the DEDS for the duplication analysis. This coming year should see us fully into the program with resultant benefits which we hope to be able to furnish you at the next annual Data Element Symposium by the Bureau of Standards.

APPENDIX A

Second National Symposium on the Management of Data Elements in Information Processing

Symposium Committee

Mr. David V. Savidge PROGRAM CHAIRMAN	Computer Network Corp. (COMNET), 5185 MacArthur Blvd., Wash., D.C. 20016 (202)244-1900
Mrs. Hazel E. McEwen ARRANGEMENTS CHAIRMAN	National Bureau of Standards, Technology Bldg., Room B226. Wash., D.C. 20234 (301)921-3157
Mr. Harry S. White, Jr.	National Bureau of Standards, Technology Bldg., Room B226. Wash., D.C. 20234 (301)921-3157

Session Chairmen

Mr. W. Scott Haynie	Western Union, 90 McKee Drive, Mahwah, New Jersey 07430 (201)529-4600/x-2126
Mr. William H. Kenworthey, Jr.	National Security Agency (ESS), Room 9A187-1, Fort George Meade, Maryland 20755 (301)688-7053/4
Miss Sheila M. Smythe[1]	Blue Cross-Blue Shield, 622 Third Ave., New York, New York 10017 (212)490-5421
Dr. William M. Taggart, Jr.	Florida International Univ., Tamiami Trail, Miami, Florida 33199 (305)552-2675

Guest Speakers

Mr. Robert W. Bemer	Honeywell Information Systems, Room B106, P.O. Box 6000, Phoenix, Arizona 85005 (602)993-2569
Dr. Ruth M. Davis	National Bureau of Standards, Administration Bldg., Room A200, Wash., D.C. 20234 (301)921-3151
Prof. Martha E. Williams[2]	Univ. of Illinois, Coordinated Science Lab (5-101), Urbana, Illinois 61801 (217)333-1000

Presenters and Authors of Papers

Mr. Irving Alderman	U.S. Army Research Institute for the Behavioral & Social Sciences, 1300 Wilson Blvd., Arlington, Virginia 22209 (202)694-3645
Dr. Donald R. Arnold	Dept. of Higher Education, State of New Jersey, 225 West State Street, P.O. Box 1293, Trenton, New Jersey 08625 (614)421-6940

---

[1] Cyril Brosnan, Blue Cross-Blue Shield, substituted.

[2] Co-authors-Scott E. Preece and Sandra H. Rouse.

Mr. George W. Covill	Vitro Labs Div., Automation Industries, Inc., 14000 Georgia Avenue, Silver Spring, Maryland 20910 (301)871-4502
Dr. Dennis Dance	Univ. of Arkansas at Little Rock, UALR, 33rd & Univ., Little Rock, Arkansas 72204 (501)568-2200
Mr. Paul-Andre Desjardins	Hospital Saint-Michel-Archange, Mastai, Quebec, P.Q., CANADA G1J 2G3 (418)661-7781
Mr. L. D. England[3]	Texas State Dept. of Public Welfare, 211 East Riverside Drive, Winters Bldg., Austin, Texas 78704 (512)475-6531
Mr. Richard H. Fahnline	U.S. Civil Service Commission, 925 25th Street, N.W. - #401, Wash., D.C. 20037 (202)337-0272
Dr. E. R. Gabrieli	State Univ. of New York at Buffalo & E. J. Meyer Memorial Hospital, 462 Grider Street, Buffalo, New York 14215 (716)894-1212
Mr. Eli Hellerman	Bureau of the Census, Room 1065-3, Suitland, Maryland 20233 (301)763-5680
Mr. Aaron Hochman	Office of the Asst. Secretary of Defense, (Installations & Logistics), Room 7S69, Hoffman Bldg. 2, 200 Stovall Str., Alexandria, Virginia 22332 (202)325-8814 or 9324
Mr. Michael A. Huffenberger	Chemical Abstracts Service, The Ohio State Univ., Columbus, Ohio 43210 (614)421-6940
Mr. Victor G. Kehler	Plans & Programs Div., Directorate of Admin., HQs U.S. Air Force, Wash., D.C. 20330 (202)756-2371
Mr. Richard J. Kirkbride	Natl. Military Command System Support Center, Military Studies & Analysis Directorate, Logistics Data Div., Wash., D.C. 20301 (202)697-3429
Dr. John R. Kraska	The Upjohn Company, Code 9960-41-0, Kalamazoo, Michigan 49001 (616)382-4000/x-3094
Mr. Robert M. Landau	Science Information Association, 3514 Plyers Mill Road, Kensington, Maryland 20795 (301)949-0220
Mr. John T. Langan	Distribution Codes, Inc., 401 Wythe Street, Alexandria, Virginia 22314 (703)683-6444
Dr. Edward Y. S. Lee and E.K.C. Lee	MTS, Jet Propulsion Lab, Calif. Institute of Technology, 4800 Oak Grove Drive, Pasadena, Calif. 91103 (213)354-4321
Mr. J. R. Nelson	The Upjohn Company, Kalamazoo, Michigan 49001 (616)382-4000

---

[3] Co-authors--S. L. Eberle, B. H. Schiff, and A. S. Huffman

Dr. Udo W. Pooch	Texas A&M Univ., Industrial Engr. Dept., College Station, Texas 77843
Mr. Fred Puente	Automated Logistics Mgmt. Systems Agency (AMXAL-MBD), U.S. Army Materiel Command (AMC), P.O. Box 1578, St. Louis, Missouri 63188 (314)268-6001
Mr. John Roberts	New York State Dept. of Motor Vehicles, Empire State Plaza, Albany, New York 12228 (518)482-8212
Mr. Curg Shields	M. Bryce & Associates, Inc., 1248 Springfield Pike, Cincinnati, Ohio 45215 (513)761-8400
Dr. William M. Taggart, Jr.	Florida International Univ., Tamiami Trail, Miami, Florida 33199 (305)552-2675
Mr. V. N. Vaughan, Jr.	A.T.&T. Company, 195 Broadway, New York, New York 10007 (212)393-3955
Mr. Marvin G. Wallis	NASA, Information Systems Office, Code YMC, Washington, D.C. 20546 (202)755-3122
Mr. Harry S. White, Jr.	National Bureau of Standards, Technology Bldg., Room B226, Wash., D.C. 20234 (301)921-3157
Ms. Edith F. Young	HQs U.S. Army Materiel Command (AMCMS-IR), 5001 Eisenhower Avenue, Alexandria, Virginia 22333 (202)274-9051/52

FUTURE DATA ELEMENT MANAGEMENT CONFERENCES

TO: Harry S. White, Jr.
Associate Director for ADP Standards
Institute for Computer Sciences and
Technology
National Bureau of Standards
Washington, D.C. 20234

FROM: (name, address, and telephone number)

___ Please add my name to your mailing list for announcements and information relating to future data element management conferences.

___ I wish to present a paper on the following subject: ____________________

____

____

___ I would like to participate on a panel or in an open forum on the following subject(s): ______________________________

____

____

I suggest the following subjects or subject areas for consideration at the future conferences: ______________________________

____

____

Other recommendations: ______________________________

____

____

____

____

NBS-114A (REV. 7-73)

U.S. DEPT. OF COMM. **BIBLIOGRAPHIC DATA SHEET**	1. PUBLICATION OR REPORT NO. PB 249530	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE

Management of Data Elements in Information Processing
(Proceedings of the Second National Symposium,
1975 October 23-24)

5. Publication Date
1976 April

6. Performing Organization Code

7. AUTHOR(S) Various
Editor: Hazel E. McEwen

8. Performing Organ. Report No.
NBSIR 76-1015

9. PERFORMING ORGANIZATION NAME AND ADDRESS

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

10. Project/Task/Work Unit No.
6009580

11. Contract/Grant No.

12. Sponsoring Organization Name and Complete Address *(Street, City, State, ZIP)*

American National Standards Institute Committee X3L8
and the National Bureau of Standards.

13. Type of Report & Period Covered
Final

14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES

16. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

Continuing technological advances in computers and communications make possible the integration of data systems and the exchange of data among them on an expanding scale. However, the full effect of these advances cannot be realized unless the need for uniform understanding of the common information (data elements) and their expression in data systems is recognized and a means provided to effectively manage this information. The increasing interrelationships among the data systems of Federal, State and local governments, and with industry and the public add emphasis and dimension to the need for the improved management of data elements in information processing.

These Proceedings are for the second Symposium on the Management of Data Elements in Information Processing held at the National Bureau of Standards on 1975 October 23-24. Over 300 representatives of Federal and State governments, industry and universities from 29 states, from Japan, and the United Kingdom were in attendance. Twenty-nine speakers discussed the role of the data manager, communications needs for data standards, data element directories, standard codes for character and control, use of check characters, data elements in bibliographic data bases, product coding, coding for clinical medicine, human factors, data resource management, data base management systems, and other subjects related to data standardization and data management efforts.

17. KEY WORDS *(six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons)*

American National Standards; American National Standards Institute; data; data base systems; data elements; data management; data processing; Federal Information Processing Standards; information interchange; information processing; information systems.

18. AVAILABILITY ☐ Unlimited

☐ For Official Distribution. Do Not Release to NTIS

☐ Order From Sup. of Doc., U.S. Government Printing Office
Washington, D.C. 204 , SD Cat. No. C13

☒ Order From National Technical Information Service (NTIS)
Springfield, Virginia 22151

19. SECURITY CLASS (THIS REPORT)
UNCLASSIFIED

21. NO. OF PAGES
278

20. SECURITY CLASS (THIS PAGE)
UNCLASSIFIED

22. Price $9.25 Paper Copy; $2.25 Microfiche Copy.

USCOMM-DC 29042-P74

U.S. GOVERNMENT PRINTING OFFICE: 19